OXFORD MASTER SERIES IN PARTICLE PHYSICS,
ASTROPHYSICS, AND COSMOLOGY

SECOND EDITION

Particle Astrophysics

Donald Perkins

OXFORD MASTER SERIES IN PARTICLE PHYSICS, ASTROPHYSICS, AND COSMOLOGY

OXFORD MASTER SERIES IN PHYSICS

The Oxford Master Series is designed for final year undergraduate and beginning graduate students in physics and related disciplines. It has been driven by a perceived gap in the literature today. While basic undergraduate physics texts often show little or no connection with the huge explosion of research over the last two decades, more advanced and specialized texts tend to be rather daunting for students. In this series, all topics and their consequences are treated at a simple level, while pointers to recent developments are provided at various stages. The emphasis in on clear physical principles like symmetry, quantum mechanics, and electromagnetism which underlie the whole of physics. At the same time, the subjects are related to real measurements and to the experimental techniques and devices currently used by physicists in academe and industry. Books in this series are written as course books, and include ample tutorial material, examples, illustrations, revision points, and problem sets. They can likewise be used as preparation for students starting a doctorate in physics and related fields, or for recent graduates starting research in one of these fields in industry.

CONDENSED MATTER PHYSICS

1. M.T. Dove: *Structure and dynamics: an atomic view of materials*
2. J. Singleton: *Band theory and electronic properties of solids*
3. A.M. Fox: *Optical properties of solids*
4. S.J. Blundell: *Magnetism in condensed matter*
5. J.F. Annett: *Superconductivity, superfluids, and condensates*
6. R.A.L. Jones: *Soft condensed matter*

17. S. Tautz: *Surfaces of condensed matter*
18. H. Bruus: *Theoretical microfluidics*

ATOMIC, OPTICAL, AND LASER PHYSICS

7. C.J. Foot: *Atomic physics*
8. G.A. Brooker: *Modern classical optics*
9. S.M. Hooker, C.E. Webb: *Laser physics*

15. A.M. Fox: *Quantum optics: an introduction*
16. S.M. Barnett: *Quantum information*

PARTICLE PHYSICS, ASTROPHYSICS, AND COSMOLOGY

10. D.H. Perkins: *Particle astrophysics, second edition*
11. Ta-Pei Cheng: *Relativity, gravitation and cosmology*

STATISTICAL, COMPUTATIONAL, AND THEORETICAL PHYSICS

12. M. Maggiore: *A modern introduction to quantum field theory*
13. W. Krauth: *Statistical mechanics: algorithms and computations*
14. J.P. Sethna: *Statistical mechanics: entropy, order parameters, and complexity*

Particle Astrophysics

Second Edition

D. H. PERKINS

Particle and Astrophysics Department
Oxford University

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries

Published in the United States
by Oxford University Press Inc., New York

© Oxford University Press 2003, 2009

The moral rights of the author have been asserted
Database right Oxford University Press (maker)

First Edition 2003
Second Edition 2009

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system, or transmitted, in any form or by any means,
without the prior permission in writing of Oxford University Press,
or as expressly permitted by law, or under terms agreed with the appropriate
reprographics rights organization. Enquiries concerning reproduction
outside the scope of the above should be sent to the Rights Department,
Oxford University Press, at the address above

You must not circulate this book in any other binding or cover
and you must impose the same condition on any acquirer

British Library Cataloguing in Publication Data

Data available

Library of Congress Cataloging in Publication Data

Data available

Typeset by Newgen Imaging Systems (P) Ltd., Chennai, India
Printed in Great Britain
on acid-free paper by
CPI Antony Rowe, Chippenham, Wiltshire

ISBN: 978–0–19–954545–2
ISBN: 978–0–19–954546–9

10 9 8 7 6 5 4 3 2 1

To my family
For their patience and encouragement

This page intentionally left blank

Preface to Second Edition

This is an enlarged and updated version of the first edition published in 2003. In a rapidly evolving field, emphasis has of course been placed on the most recent developments. However, I have also taken the opportunity to re-arrange the material and present it in more detail and at somewhat greater length.

For convenience, the text has been divided into three parts. Part 1, containing Chapters 1–4, deals basically with the fundamental particles and their interactions, as observed in laboratory experiments, which are covered by the so-called Standard Model of particle physics. This model gives an extremely exact and detailed account of an immense mass of experimental data obtained at accelerators worldwide, although some postulated phenomena such as the Higgs boson have still to be observed. Developments beyond the original Standard Model, particularly the subject of neutrino masses and flavour oscillations, are included, as well as possible extensions of the model, such as supersymmetry and the grand unification of the fundamental interactions. I have also taken the opportunity to present in Chapter 2, a short account of relativistic transformations, the equivalence principle and solutions of the field equations of general relativity which are important for astrophysics.

Part 2 (Chapters 5–8) describes the present picture of the cosmos in the large, with emphasis on the basic parameters of the early universe, which are now becoming more accurately known and expressed in the so-called Concordance Model of cosmology. This part also underlines the great questions and mysteries in cosmology: the nature of dark matter; the nature of dark energy and the magnitude of the cosmological constant; the matter–antimatter asymmetry of the universe; the precise mechanism of inflation; and, just as is the case for the 20 or so parameters describing the Standard Model of particles, the arbitrary nature of the parameters in the Concordance Model.

Part 3 (Chapters 9 and 10) is concerned with the study of the particles and radiation which bombard us from outer space, and to the stellar phenomena, such as pulsars, active galactic nuclei, black holes, and supernovae which appear to be responsible for this 'cosmic rain'. We encounter here some of the most energetic and bizarre processes in the universe, with new experimental discoveries being made on an almost daily basis.

By and large, the above division of the subject matter in a sense also reflects the state of our knowledge in the three cases. One could say that particle physics at accelerators in Part 1 is an extremely well-understood subject, with agreement between theory and experiment better than one part per million in the case of quantum electrodynamics. Whatever the form might be of an ultimate 'theory of everything'—-if there ever is one—-the Standard Model of particle physics will surely be part of it, even if it only accounts for a paltry 4% of the energy density of the universe at large.

Our knowledge of the basic parameters of cosmology in Part 2, while less exact is now, as compared with just a decade ago, reaching quite remarkable levels of precision, as described in the Concordance Model. In contrast, the wide-ranging study of the particles and radiation in Part 3 leaves very many open questions and is probably the least well-understood aspect of particle astrophysics. For example, a century after they were first discovered, it is only recently that we have gained some idea on how the cosmic rays are accelerated to the very highest energies (of the order of 10^{20} eV) that can be detected and, more than 30 years after their first detection, we still do not know what is the underlying mechanism of γ-ray bursts, perhaps the most violent events taking place in the present universe.

Some subjects appropriate to particle astrophysics have been left out, either through lack of space or because I thought they might be too advanced or too speculative. As in the first edition, the general theory of relativity has been omitted, although in Chapter 2, I have tried to give some plausibility arguments, based on the equivalence principle and special relativity, to illustrate important solutions of the Einstein field equations. General relativity is in any case adequately covered by the companion volume on *Gravitation, Relativity and Cosmology* by T.P. Cheng in the Oxford Master Series.

As for the first edition, the text is intended for physics undergraduates in their third or later years, so I have kept the presentation and mathematical treatment at a reasonable level. I believed that it was more important to concentrate on the outstanding developments and the burning questions in a very exciting and very wide-ranging subject, rather than spend time and space on long theoretical discussion. At no point have I hesitated to sacrifice mathematical rigour for the sake of brevity and clarity. Again, I have sprinkled a few worked examples throughout the text, which is supplemented with sets of problems at the end of chapters, with answers and some worked solutions at the end of the book.

Acknowledgements

For permission to reproduce photographs, figures and diagrams I am indebted to the authors cited in the text and to the following individuals, laboratories, journals and publishers:

Addison Wesley Publishers for Figure 5.9, from *The Early Universe*, by E.W. Kolb and M.S. Turner (1990); for Figure 10.1, from *Introduction to Nuclear Physics*, by H.A. Enge (1972); for Figures 1.12, 1.13, 1.14, 5.4, 5.7, 5.10, 9.8 and 9.9 from *Introduction to High Energy Physics*, 3rd edition, by D.H. Perkins;
American Physical Society, publishers of *Physical Review*, for Figure 1.2, and publishers of *Physical Review Letters*, for Figure 7.12;
Astronomy and Astrophysics, for Figure 7.2;
Astrophysical Journal, for Figures 5.2, 7.13, 7.14 and 7.6;
Annual Reviews of Nuclear and Particle Science for Figure 9.2;
Elsevier Science, publishers of *Physics Reports*, for Figure 9.13, and publishers of *Nuclear Physics B*, for Figure 10.3;

Institute of Physics, publishers of *Reports on Progress in Physics*, for Figure 1.7;
CERN Information Services, Geneva for Figures 1.1 and 1.6;
DESY Laboratory, Hamburg, for Figure 1.5;
European Southern Observatory, for Figure 9.6;
Fermilab Visual Media Services, Fermilab, Chicago for Figure 1.11;
Prof. D. Clowe for Figure 7.6; the SNO collaboration for Figure 9.25; Prof. M. Tegmark for Figure 8.7; Prof. A.G. Riess for Figure 7.13 and 7.14; Prof. Chris Carilli, NARO, for Figure 9.20; Prof. Y. Suzuki and Prof. Y. Totsuka of the Superkamiokande Collaboration for Figure 9.22 and Figure 4.7; Prof. Trevor Weekes, Whipple Observatory, Arizona for Figure 9.11.

This page intentionally left blank

Contents

Part 1 Particles and Interactions

1 Quarks and leptons and their interactions 3
1.1 Preamble 3
1.2 Quarks and leptons 4
1.3 Fermions and bosons: the spin-statistics theorem; supersymmetry 9
1.4 Antiparticles 9
1.5 The fundamental interactions: boson exchange 11
1.6 The boson couplings to fermions 14
1.7 The quark–gluon plasma 21
1.8 The interaction cross section 21
1.9 Examples of elementary particle cross sections 24
1.10 Decays and resonances 30
1.11 Examples of resonances 32
1.12 New particles 34
1.13 Summary 35
Problems 36

2 Relativistic transformations and the equivalence principle 39
2.1 Coordinate transformations in special relativity 39
2.2 Invariant intervals and four-vectors 41
2.3 The equivalence principle: clocks in gravitational fields 42
2.4 General relativity 47
2.5 The Schwarzschild line element, Schwarzschild radius, and black holes 49
2.6 The gravitational deflection of light by a point mass 51
2.7 Shapiro time delay 52
2.8 Orbital precession 53
2.9 The Robertson–Walker line element 54
2.10 Modifications to Newtonian gravity? 55
2.11 Relativistic kinematics: four-momentum; the Doppler effect 56
2.12 Fixed-target and colliding-beam accelerators 57
Problems 59

3 Conservation rules, symmetries, and the Standard Model of particle physics 60

3.1 Transformations and the Euler–Lagrange equation 60
3.2 Rotations 62
3.3 The parity operation 62
3.4 Parity conservation and intrinsic parity 63
3.5 Parity violation in weak interactions 65
3.6 Helicity and helicity conservation 67
3.7 Charge conjugation invariance 69
3.8 Gauge transformations and gauge invariance 69
3.9 Superstrings 73
3.10 Gauge invariance in the electroweak theory 74
3.11 The Higgs mechanism of spontaneous symmetry breaking 75
3.12 Running couplings 77
3.13 Vacuum structure in gauge theories 83
3.14 CPT theorem and CP and T symmetry 83
3.15 CP violation in neutral kaon decay 84
3.16 CP violation in the Standard Model: the CKM matrix 87
3.17 Summary 89
Problems 90

4 Extensions of the Standard Model 92

4.1 Neutrinoless double beta decay 92
4.2 Neutrino masses and flavour oscillations 94
4.3 Grand unified theories: proton decay 97
4.4 Grand unification and the neutrino see-saw mechanism 100
4.5 Hierarchies and supersymmetry 102
4.6 Summary 103
Problems 104

Part 2 The Early Universe

5 The expanding universe 107

5.1 The Hubble expansion 107
5.2 Olbers' paradox 113
5.3 The Friedmann equation 114
5.4 The sources of energy density 117
5.5 Observed energy densities 120
5.6 The age and size of the universe 123
5.7 The deceleration parameter 126
5.8 Cosmic microwave background radiation (CMB) 127
5.9 Anisotropies in the microwave radiation 130
5.10 Particles and radiations in the early universe 131
5.11 Photon and neutrino densities 134
5.12 Radiation and matter eras 135

5.13 The eras of matter–radiation equality 138
5.14 Summary 139
Problems 140

6 Nucleosynthesis and baryogenesis 142
6.1 Primordial nucleosynthesis 142
6.2 Baryogenesis and matter–antimatter asymmetry 146
6.3 The baryon–photon ratio in the Big Bang 148
6.4 The Sakharov criteria 150
6.5 The baryon–antibaryon asymmetry: possible scenarios 151
6.6 Summary 154
Problems 155

7 Dark matter and dark energy components 156
7.1 Preamble 156
7.2 Dark matter in galaxies and clusters 157
7.3 Gravitational lensing 159
7.4 Evidence for dark matter from gravitational lensing 160
7.5 Microlensing and MACHOs 163
7.6 The lensing probability: optical depth 165
7.7 Baryonic dark matter 166
7.8 Neutrinos 167
7.9 Axions 168
7.10 Axion-like particles 169
7.11 Weakly interacting massive particles 170
7.12 Expected WIMP cross-sections and event rates 173
7.13 Experimental WIMP searches 174
7.14 Dark energy: high redshift supernovae and Hubble plot at large z 176
7.15 Vacuum energy: the Casimir effect 182
7.16 Problems with the cosmological constant and dark energy 184
7.17 Summary 186
Problems 187

8 Development of structure in the early universe 188
8.1 Preamble 188
8.2 Galactic and intergalactic magnetic fields 189
8.3 Horizon and flatness problems 190
8.4 Inflation 192
8.5 Chaotic inflation 196
8.6 Quantum fluctuations and inflation 198
8.7 The spectrum of primordial fluctuations 200
8.8 Gravitational collapse and the Jeans mass 202
8.9 The growth of structure in an expanding universe 205
8.10 Evolution of fluctuations during the radiation era 206

8.11 Cosmological limits on neutrino mass from fluctuation spectrum 210
8.12 Growth of fluctuations in the matter-dominated era 212
8.13 Temperature fluctuations and anisotropies in the CMB 213
8.14 The Angular spectrum of anisotropies: ‘acoustic peaks’ in the distribution 216
8.15 Experimental observation and interpretation of CMB anisotropies 222
8.16 Polarization of the cosmic microwave radiation 223
8.17 Summary 224
Problems 226

Part 3 Particles and Radiation in the Cosmos

9 Cosmic particles 229
9.1 Preamble 229
9.2 The composition and spectrum of cosmic rays 230
9.3 Geomagnetic and solar effects 233
9.4 Acceleration of cosmic rays 237
9.5 Secondary cosmic radiation: pions and muons 239
9.6 Passage of charged particles and radiation through matter 240
9.7 Development of an electromagnetic cascade 243
9.8 Extensive air showers: nucleon- and photon-induced showers 245
9.9 Detection of extensive air showers 245
9.10 Point sources of γ-rays 247
9.11 γ-Ray bursts 249
9.12 Ultra-high-energy cosmic ray showers: the GZK cut-off 251
9.13 Point sources of ultra high energy cosmic rays 253
9.14 Radio galaxies and quasars 253
9.15 Atmospheric neutrinos: neutrino oscillations 257
9.16 Solar neutrinos 260
9.17 Neutrino oscillations in matter 263
9.18 Point sources of high-energy neutrinos 263
9.19 Gravitational radiation 264
9.20 The binary pulsar 267
9.21 Detection of gravitational waves 269
9.22 Summary 270
Problems 271

10 Particle physics in stars and galaxies 273
10.1 Preamble 273
10.2 Stellar evolution—the early stages 273
10.3 Hydrogen burning: the p-p cycle in the Sun 276

10.4 Helium burning and the production of carbon and oxygen 278
10.5 Production of heavy elements 280
10.6 Electron degeneracy pressure and stellar stability 281
10.7 White dwarf stars 284
10.8 Stellar collapse: type II supernovae 285
10.9 Neutrinos from SN 1987A 288
10.10 Neutron stars and pulsars 291
10.11 Black holes 294
10.12 Hawking radiation from black holes 295
10.13 Summary 297
Problems 298

A Table of physical constants 299

B Yukawa theory and the boson propagator 301

C Perturbative growth of structure in the early universe 303
C.1 Growth in the matter-dominated era 306

D The MSW mechanism in solar neutrino interactions 308

Answers to problems 312

References 329

Bibliography 333

Index 335

This page intentionally left blank

Part 1

Particles and Interactions

This page intentionally left blank

Quarks and leptons and their interactions

1

1.1 Preamble

High-energy particle physics is concerned with the study of the fundamental constituents of matter and the interactions between them. Experiments in this field have been carried out with giant accelerators and their associated detection equipment, which have probed the structure of matter down to very small scales, of order 10^{-17} m, that is, about one hundredth of the radius of a proton. In contrast, astrophysics is concerned with the structure and evolution of the universe in the large, including the study of the behaviour of matter and radiation on enormous scales, up to around 10^{26} m. The experimental observations have been made with telescopes on the Earth or on satellites, covering the visible, infrared, and ultraviolet regions of the spectrum, as well as with detectors of radio waves, X-rays, γ-rays, and neutrinos. These have revealed an astonishing range of extra-terrestrial phenomena from the most distant regions of the cosmos.

In describing this scientific adventure into what is, even today, still very largely unknown territory, the object of this text has been to show how studies of particles on a laboratory scale have helped in our understanding of the development of the universe, and conversely how celestial observations have, in turn, shed light on our understanding of particle interactions. Although in this chapter we discuss the constitution of matter as determined by accelerator experiments on Earth, it is becoming clear that on cosmological scales, other quite different forms of matter and energy may be important or even dominant, as will be discussed in the later chapters. Nevertheless, it is clear that a thorough understanding of the properties and interactions of elementary particles in laboratory experiments is essential in the discussion of astrophysical phenomena on the grandest scales.

First we should note the units employed in the study of the fundamental quark and lepton constituents of matter. The unit of length is the femtometer (1 fm $= 10^{-15}$ m), an appropriate unit because, for example, the charge radius of a composite particle such as a proton is 0.8 fm. The typical energy scale is the giga electron volt (1 GeV $= 10^9$ eV); for example, the mass energy equivalent of a proton is $M_\mathrm{p}c^2 = 0.938$ GeV. Table 1.1 lists the units employed in high-energy physics, together with their equivalents in SI units. A list of appropriate physical constants is given in Appendix A.

In the description of particle interactions at the quantum level, the quantities $\hbar = h/2\pi$ and c frequently occur, and it is convenient to employ so-called natural units, which set $\hbar = c = 1$. Having chosen these two units, we are free to specify just one more unit, which is taken as that of energy, the GeV =

Table 1.1 Units in high-energy physics

Quantity	High-energy unit	Value in SI units
Length	1 fm	10^{-15} m
Energy	1 GeV	1.602×10^{-10} J
Mass, E/c^2	1 GeV/c^2	1.78×10^{-27} kg
$\hbar = h/2\pi$	6.588×10^{-25} GeV s	1.055×10^{-34} J s
c	2.998×10^{23} fm s^{-1}	2.998×10^{8} m s^{-1}
$\hbar c$	0.1975 GeV fm	3.162×10^{-26} J m

Table 1.2 Quark and lepton flavours

Symbol	Name	Q/|e|	Symbol	Name	Q/|e|
u	up	+2/3	e	electron	−1
d	down	−1/3	ν_e	e-neutrino	0
c	charm	+2/3	μ	muon	−1
s	strange	−1/3	ν_μ	μ-neutrino	0
t	top	+2/3	τ	tauon	−1
b	bottom	−1/3	ν_τ	τ-neutrino	0

10^9 eV (the giga electron volt). The unit of mass is then $Mc^2/c^2 = 1$ GeV, that of length is $\hbar c/Mc^2 = 1\ \text{GeV}^{-1} = 0.1975$ fm, and that of time is $\hbar c/Mc^3 = 1\ \text{GeV}^{-1} = 6.59 \times 10^{-25}$ s.

1.2 Quarks and leptons

In the so-called Standard Model of particle physics, which is strongly supported by extensive laboratory experiments and is more fully discussed in Chapter 3, the material universe is assumed to be built from a small number of fundamental constituents, the *quarks* and the *leptons*. The names of these, together with their electrical charges are given in Table 1.2. All these particles are *fermions*, that is, they have half-integral intrinsic angular momentum or spin, $1/2\hbar$. For each of the particles in the table there is an *antiparticle*, with the opposite value of electric charge and magnetic moment, but with identical mass and lifetime to those of the particle. For example the positron (see Fig. 1.2) e^+ is the antiparticle of the electron, e^-. In contrast with the proton and neutron, which are extremely small but nevertheless extended objects, the quarks and leptons are considered to be *pointlike*: as far as we know today, they are truly elementary and are not composed of other, even more fundamental entities.

Considering first the charged leptons, the electron is familiar to everyone. The muon μ and the tauon τ are heavier, highly unstable versions of the electron, with mean lifetimes of 2.2×10^{-6} s and 2.9×10^{-13} s respectively. The properties of the charged leptons—their masses, lifetimes, magnetic moments—are very well measured. In particular, magnetic moments are in very precise agreement with the predictions of quantum electrodynamics, at the level of one part per million or better. For the neutral leptons—neutrinos—matters are much more complicated and less well understood.

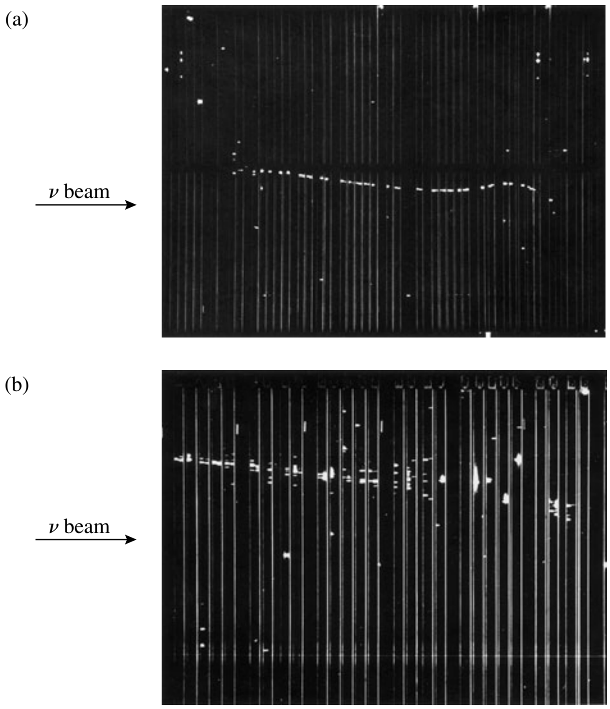


Fig. 1.1 Interaction of neutrino beam, from left, of about 1 GeV energy, in a CERN experiment employing a spark chamber detector. This consists of an array of parallel, vertical metal plates maintained at high voltages. A charged particle will ionise the gas between the plates, and this leads to a complete breakdown of the gas in a spark (Geiger) discharge. Thus charged particle trajectories appear as rows of sparks. The event at the top is attributed to a muon-type neutrino. Upon interaction in the plate it transforms to a muon, which traverses many plates before coming to rest. The event at the bottom is due to an electron-type neutrino, transforming to an electron. The latter generates scattered sparks characteristic of an electron–photon shower, as described in Chapter 9, quite distinct from the rectilinear muon track. In both cases, the reactions are 'elastic', of the form $\nu_l + n \rightarrow l + \mathrm{p}$, where $l = \mu$ or e, and the recoiling proton is stopped inside the plate. (Courtesy CERN Information Services).

1.2.1 Neutrinos

Associated with each charged lepton is a neutral lepton, called a neutrino, denoted by the generic symbol ν. A different neutrino ν_e, ν_μ, or ν_τ is associated with each different type or *flavour* of charged lepton. For example, in nuclear beta decay, a (bound) proton in a nucleus transforms to a neutron together with a positron e^+ which is emitted together with an electron-type neutrino, that is, $\mathrm{p} \rightarrow \mathrm{n} + \mathrm{e}^+ + \nu_\mathrm{e}$. In a subsequent interaction, this neutrino, if it is energetic enough, may transform into an electron, that is, $\nu_\mathrm{e} + \mathrm{n} \rightarrow \mathrm{e}^- + \mathrm{p}$, but not into a charged muon or tauon (see Fig. 1.1 for examples of such transformations). In any interaction, the flavour is conserved.

All the particles (and their antiparticles) in Table 1.2, with the exception of the neutrinos, are fermions with two spin substates each: relative to the momentum (z-) axis, the spin components are $\mathrm{s}_z = \pm 1/2\hbar$. However, a neutrino has only one spin state, $\mathrm{s}_z = -1/2\hbar$, while an antineutrino has $\mathrm{s}_z = +1/2\hbar$ only. The spin and momentum vectors together define a 'screw sense', the neutrino being left-handed (LH) and the antineutrino being right-handed (RH)—see Section 3.6. Why there is such a left-right asymmetry, with the interaction cross-section in matter of antineutrinos different from that of neutrinos, is unknown. The story here may, however, be more complicated. The above description assumes that neutrinos and antineutrinos are so-called Dirac particles, quite distinct from each other, just like the charged leptons and antileptons. A neutrino that can assume only one of the two possible spin substates must have velocity $v = c$ and be massless (otherwise, in a transformation to another reference frame, necessarily moving at $v < c$, it would be possible to reverse the direction of spin relative to the momentum). However, simply because neutrinos are uncharged, the other possibility is that neutrinos are *their own antiparticles*. These so-called 'Majorana' neutrinos occur in spin-up and spin-down substates, labelled 'neutrino' and 'antineutrino' in the Dirac picture. Unfortunately, experiments at the present time are unable to differentiate between these two prescriptions. In Chapters 4 and 6 we describe how massive Majorana neutrinos have been invoked in a mechanism to account for the baryon–antibaryon asymmetry in the universe, which is one of the big puzzles in cosmology. However, because of common usage, and except when we specifically discuss Majorana particles, we will speak of neutrinos and antineutrinos as particle and antiparticle.

To further complicate matters, it turns out that, while the charged leptons are described by wavefunctions which are unique mass eigenstates, the neutrinos are not, but superpositions of mass eigenstates, with slightly different mass values and, for a given momentum therefore, slightly different velocities. As a consequence, in travelling through empty space, phase differences develop as the different mass eigenstates get 'out of step', appearing as *oscillations* in the neutrino flavour between ν_e, ν_μ, and ν_τ. Neutrino oscillations are discussed fully in Chapters 4 and 9; they measure the mass differences between the eigenstates, and these are tiny, of order 0.1 eV/c^2. The corresponding wavelengths of the flavour oscillations are extremely long—100's or 1000's of kilometres for neutrino energies of order 1 GeV. Neutrino oscillations had actually been proposed in 1964, but because of these tiny mass differences, it was to take some 30 years of research—and 60 years after Pauli first postulated the neutrino – to discover them. Direct measurements of neutrino mass (rather than mass differences) shown in Table 1.3, obtained, for example, in the case of ν_e from the shape of the end-point in tritium beta decay, give upper limits which at present are much larger.

The neutrinos turn out to be of great importance in the cosmology of the early universe, as discussed later. Next to the microwave photons constituting the relic electromagnetic radiation from the Big Bang (411 per cm^3 throughout space), relic neutrinos are by far the most abundant particles in the universe. Their number density of 340 per cm^3 is to be compared with only about 1.25×10^{-7} protons, neutrons, or electrons per cm^3. As discussed in detail in Chapter 8, it seems that in the primordial universe during its first 400,000 years, before radiation decoupled from matter, neutrinos played an important and indeed crucial role in the development of cosmic structures on the large scales of galaxy clusters and superclusters.

1.2.2 Quark flavours

The quarks in Table 1.2 have fractional electric charges, of $+2|e|/3$ and $-|e|/3$ where $|e|$ is the numerical value of the electron charge. As for the charged leptons, the masses increase as we go down the table (see also Tables 1.3 and 1.4). Apart from charge and spin, the quarks, like the leptons, have an extra internal degree of freedom, again called the *flavour*. The odd names for the various quark flavours—'up', 'down', 'charm', etc.—have arisen historically. Just as for the leptons, the six flavours of quark are arranged in three doublets, the components of which differ by one unit of electric charge.

While the leptons exist as free particles, the quarks do not (but see the remarks later on the quark–gluon plasma). It is a peculiarity of the strong force between the quarks that, at normal energies, they are always found associated in quark composites called *hadrons*. These are of two types: *baryons* consist of three quarks, QQQ, while *mesons* consist of a quark–antiquark pair, $Q\bar{Q}$. For example,

$$\begin{array}{llll} \text{Proton} & = u\,u\,d & \text{Neutron} & = d\,d\,u \\ \text{Pion}\,\pi^+ & = u\,\bar{d} & \pi^- & = \bar{u}\,d \end{array}$$

The common material of the world today is built from u and d quarks, forming the protons and neutrons of atomic nuclei, which together with the electrons e^- form atoms and molecules. The heavier quarks c, s, t, and b are also observed to form baryon composites such as *sud, sdc*, ... and mesons such as $b\bar{b}$, $c\bar{c}$, $c\bar{b}$, ..., but these heavy hadrons are all highly unstable and decay rapidly to states containing u and d quarks only. Likewise the heavier charged leptons μ and τ decay to electrons and neutrinos. These heavy quarks and leptons can be produced in collisions at laboratory accelerators, or naturally in the atmosphere as a result of collisions of high-energy cosmic rays. However, they appear to play only a minor role in today's relatively cold universe. For example, while several hundred high-energy muons (coming down to earth as secondary components of the cosmic rays) pass through everyone each minute, this is a trivially small number compared with the human tally of electrons, of order 10^{28}. The cosmic ray muons add to the natural levels of radiation, coming from radioactive elements in the ground and the atmosphere, and presumably therefore they make a contribution to the natural gene mutation rate from the effects of radiation.

Of course we believe that these heavier flavours of quark and lepton would have been as prolific as the light ones at a very early, intensely hot stage of the Big Bang, when the temperature was such that the mean thermal energy kT far exceeded the mass energy of these particles. Indeed, it is clear that the type of universe we inhabit today must have depended very much, in its initial evolution, on these heavier fundamental particles—as well as on new forms of matter and energy, which (so far) cannot be produced at accelerators and the existence of which we deduce from astronomical observations.

Table 1.3 Lepton masses in energy units, mc^2

Flavour	Charged lepton mass	Neutral lepton mass
e	0.511 MeV	$\nu_e < 2.5$ eV
μ	105.66 MeV	$\nu_\mu < 0.17$ MeV
τ	1777 MeV	$\nu_\tau < 18$ MeV

Table 1.4 Constituent quark masses

Flavour	Quantum number	Approximate rest-mass, GeV/c^2
up or down	—	0.31
strange	$S = -1$	0.50
charm	$C = +1$	1.6
bottom	$B = -1$	4.6
top	$T = +1$	175

The masses of the quarks and leptons are given in Tables 1.3 and 1.4. The masses shown for neutrinos in Table 1.3 are upper limits deduced from energy and momentum conservation in decays involving neutrinos (e.g., from the kinematics of pion decay $\pi \rightarrow \mu + \nu_\mu$, or from tritium beta decay $^3\mathrm{H} \rightarrow {}^3\mathrm{He} + \mathrm{e}^- + \bar{\nu}_\mathrm{e}$). As already noted above, evidence from neutrino flavour oscillations indicates neutrino mass differences, and by inference the masses themselves, which are much less than these limits, and in the region of 0.1 eV/c^2. Upper limits of $\sim$0.5 eV/c^2 for the neutrino masses, summed over all flavours, are also obtained from analyses of the spectrum of density fluctuations of microwave radiation in the early universe, and from galaxy surveys, described in Chapter 8.

As discussed below, the quarks are held together in hadrons by the gluon carriers of the strong force, and the 'constituent' quark masses in Table 1.4 include such quark binding effects. The u and d quarks have nearly equal masses (each of about one third that of the nucleon) as indicated by the smallness of the neutron–proton mass difference of 1.3 MeV/c^2. Isospin symmetry in nuclear physics results from this near coincidence in the light quark masses.

High-energy scattering experiments often involve 'close' collisions between the quarks. In this case, the quarks can be temporarily separated from their retinue of gluons, and the so-called current quark masses which then apply are smaller than the constituent masses by about 0.30 GeV/c^2. So the current u and d quark masses are a few MeV/c^2 only. In this regard, the smallness of the neutron–proton mass difference is not such a coincidence.

In the strong interactions between the quarks, the flavour quantum number is conserved, and is denoted by the quark symbol in capitals. For example, a strange s quark has a strangeness quantum number $S = -1$, while a strange antiquark $\bar{s}$ has $S = +1$. Thus, in a collision between hadrons containing u and d quarks only, heavier quarks can be produced, but only as quark–antiquark pairs, so that the net flavour is conserved. In weak interactions, on the contrary, the quark flavour may change, for example, one can have transitions of the form $\Delta S = \pm 1$, $\Delta C = \pm 1$, etc. As an example, a baryon called the lambda hyperon of $S = -1$ decays to a proton and a pion, $\Lambda \rightarrow \mathrm{p} + \pi^-$, with a mean lifetime of 2.6×10^{-10} s, typical of a weak interaction of $\Delta S = +1$. This decay would be expressed as $sud \rightarrow uud + d\bar{u}$ in quark nomenclature.

A few words are appropriate here about the practical attainment in the laboratory of high mass scales in high-energy particle physics. The completion of Table 1.2 of fermions and Table 1.5 of bosons took over 40 years of the twentieth century, as bigger and more energetic particle accelerators were

Table 1.5 The fundamental interactions ($Mc^2 = 1$ GeV)

	Gravitational	Electromagnetic	Weak	Strong
Field boson	Graviton	Photon	W, Z	Gluon
Spin/Parity	2^+	1^-	$1^+, 1^-$	1^-
Mass	0	0	$M_W = 80.2$ GeV $M_Z = 91.2$ GeV	0
Source	mass	electric charge	weak charge	colour charge
Range, m	∞	∞	10^{-18}	$< 10^{-15}$
Coupling Constant	$\mathrm{G}M^2/4\pi\hbar c = 5 \times 10^{-40}$	$\alpha = e^2/4\pi\hbar c = 1/137$	$\mathrm{G}_F\left(Mc^2\right)^2/(\hbar c)^3 = 1.17 \times 10^{-5}$	$\alpha_s \leq 1$

able to excite production of more and more massive fundamental states. The first evidence for the existence of the lighter quarks u, d, s appeared in the 1960s, from experiments at the CERN PS (Geneva) and Brookhaven AGS (Long Island) proton synchrotrons, with beam energies of 25–30 GeV, as well as the 25 GeV electron linear accelerator at SLAC, Stanford. The weak bosons W and Z, with masses of 80 and 90 GeV/c^2 were first observed in 1983 at the CERN proton–antiproton collider with oppositely circulating beams of energy 270 GeV (see Fig. 1.6). The most massive particle so far produced, the top quark of mass 175 GeV/c^2, was first observed in 1995 at the Fermilab proton–antiproton collider (Chicago), with 900 GeV energy in each beam.

1.3 Fermions and bosons: the spin-statistics theorem; supersymmetry

As stated above, the fundamental particles consist of half-integer spin *fermions*, the quarks and leptons, the interactions of which are mediated, as described below, by integer spin *bosons.* The distinction between the two types is underlined by the *spin-statistics theorem.* This specifies the behaviour of an ensemble of identical particles, described by some wave function ψ, when any two particles, say 1 and 2, are interchanged. The probability $|\psi|^2$ cannot be altered by the interchange, since the particles are indistinguishable, so under the operation, $\psi \to \pm\psi$. The rule is as follows:

Identical bosons:	under interchange	$\psi \to +\psi$	Symmetric
Identical fermions:	under interchange	$\psi \to -\psi$	Antisymmetric

Suppose, for example, that it were possible to put two identical fermions in the *same* quantum state. Then under interchange, ψ would not change sign, since the particles are indistinguishable. However, according to the above rule ψ *must* change sign. Hence two identical fermions cannot exist in the same quantum state—the famous Pauli Principle. On the other hand, there are no restrictions on the number of identical bosons in the same quantum state, an example of this being the laser.

One important development in connection with theories unifying the fundamental interactions at very high mass scales, has been the postulate of a fermion–boson symmetry called *supersymmetry*. For every known fermion state there is assigned a boson partner, and for every boson a fermion partner. The reasons for this postulate are discussed in Chapter 3, and a list of proposed supersymmetric particles given in Table 3.2. At this point we content ourselves with the remark that if they exist, supersymmetric particles created in the early universe could be prime candidates for the mysterious *dark matter* which, as we shall see in Chapter 7, constitutes the bulk of the material universe. However, at the present time there is no direct experimental evidence for the existence of supersymmetric particles.

1.4 Antiparticles

In 1931, Dirac wrote down a wave equation describing an electron, which was first order in both space and time coordinates, and had four solutions. Two of these were for the electron, and corresponded to the two possible *spin substates* with projections $s_z = +(1/2)\hbar$ and $-(1/2)\hbar$ along the quantization axis. The other two solutions were attributed to the *antiparticle*, with similar properties to the electron except for the opposite value of the electric charge. These predictions followed from the two great conceptual advances in twentieth-century physics, namely the classical theory of relativity and the quantum mechanical description of atomic and subatomic phenomena.

The relativistic relation connecting energy E, momentum p, and rest-mass m is (see the section on relativistic kinematics in Chapter 2)

$$E^2 = p^2c^2 + m^2c^4 \tag{1.1}$$

From this equation we see that the total energy can in principle assume both negative and positive values

$$E = \pm\sqrt{p^2c^2 + m^2c^4} \tag{1.2}$$

While in classical mechanics negative energies appear to be meaningless, in quantum mechanics we represent a stream of electrons travelling along the positive x-axis by the plane wavefunction

$$\psi = A \ \exp[-i(Et - px)/\hbar] \tag{1.3}$$

where the angular frequency is $\omega = E/\hbar$, the wavenumber is $k = p/\hbar$, and A is a normalization constant. As t increases, the phase $(Et - px)$ advances in the direction of positive x. However, (1.3) can equally well represent a particle of energy $-E$ and momentum $-p$ travelling in the negative x-direction and *backwards in time*, that is, replacing Et by $(-E)(-t)$ and px by $(-p)(-x)$:

$$\begin{array}{lll} E > 0 & E < 0 & \\ t_1 \rightarrow t_2 & t_1 \leftarrow t_2 & (t_2 > t_1) \end{array}$$

A stream of negatively charged electrons flowing backwards in time is equivalent to a positive charge flowing forwards, thus with $E > 0$. Hence the negative energies are formally connected with the existence of a positive energy antiparticle e^+, the positron. This particle was first observed by Anderson in 1932, quite independently of Dirac's prediction (see Fig. 1.2). The existence of antiparticles is a general property of both fermions and bosons, but for fermions only there is a conservation rule. One can define a fermion number, +1 for a fermion and −1 for an antifermion and postulate that the total fermion number is conserved. Thus fermions can only be created or destroyed in particle–antiparticle pairs, such as e^+e^- or $Q\bar{Q}$. For example, a γ-ray, if it has energy $E > 2mc^2$ where m is the electron mass, can create a pair (in the presence of an atom to conserve momentum), and an e^+e^- pair can annihilate to γ-rays. As another example, in massive stars reaching the supernova phase, fermion number conserving reactions such as $e^+ + e^- \rightarrow \nu + \bar{\nu}$ are expected to be commonplace.

At the energies available in the laboratory, and certainly today in our relatively cold universe, leptons and baryons are strictly conserved. Thus a charged lepton (e^-, μ^-, or τ^-) is given a lepton number $L = +1$, while their antiparticles are

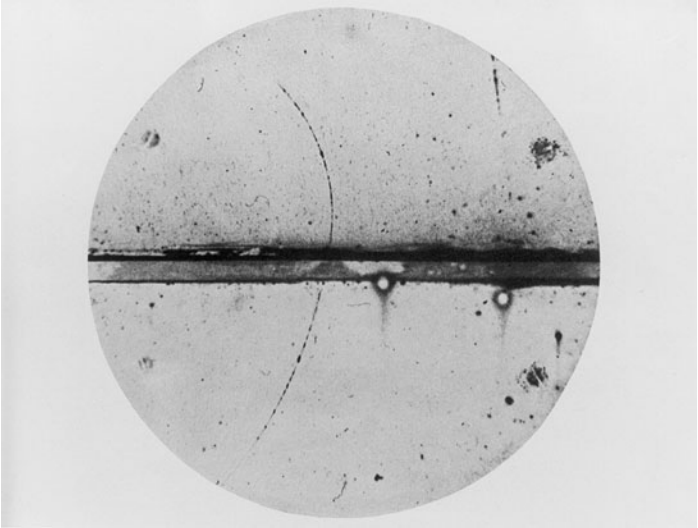


Fig. 1.2 The discovery of the positron by Anderson in 1932, in a cloud chamber experiment investigating the cosmic rays. The cloud chamber consists of a glass-fronted cylindrical tank of gas saturated with water vapour. Upon applying an expansion by means of a piston at the rear of the chamber, the gas cools adiabatically, it becomes supersaturated and water condenses as droplets, especially on charged ions created by passage of a charged particle. A magnetic field applied normal to the chamber plane allows measurement of particle momentum from track curvature. Note that this curvature is larger in the top half of the chamber, because the particle loses momentum in traversing the central metal plate. Hence it was established that the particle was positive and travelling upwards, and that its mass was very much less than that of a proton, and consistent with that of the electron. The first actual observation of electron–positron pairs was made by Blackett and Occhialini, using a counter-controlled cloud chamber in an experiment at Cambridge, contemporaneous with that of Anderson at Cal Tech. Since these early experiments, many other types of antiparticle have been observed, including the antiproton in 1955 and the anti-hydrogen atom in 1995.

assigned $L = -1$. The corresponding neutrinos ν_e, ν_μ, and ν_τ also have $L = +1$ (and antineutrinos $L = -1$), so that in a reaction such as $\nu_e + n \to e^- + p$, the total lepton number is conserved. Similarly, in this same reaction, the baryon number $B = +1$ for the neutron and proton is conserved. With three in each nucleon, quarks themselves are assigned $B = +1/3$ (and antiquarks, $B = -1/3$). As discussed more fully in Chapter 6, there are in fact no deep theoretical reasons why leptons and baryons should be conserved. Lepton and baryon conservation appears to be a low-energy phenomenon. The very pronounced baryon–antibaryon asymmetry in the universe at large suggests that in the very early, very hot universe, interactions took place which did violate lepton and baryon conservation.

1.5 The fundamental interactions: boson exchange

The elementary fermions—the quarks and leptons—are postulated to interact via the *exchange of boson mediators*, the boson carrying momentum from one fermion to the other. The rate at which momentum is exchanged in this way provides the force between the interacting particles. There are four known types of interaction, each with its characteristic boson exchange particle. They are as follows:

- The *electromagnetic* interaction occurs between all types of charged particle, and is brought about by exchange of a *photon*, with spin 1 and zero mass.
- The *strong* interactions occur between the quarks, via exchange of the *gluon*, a particle again with spin 1 and zero mass. Such interactions are responsible not only for the binding of quarks in hadrons but also for the force holding neutrons and protons together in atomic nuclei.
- The *weak* interactions take place between all types of quark and lepton. They are mediated by the exchange of *weak bosons*,$W^{\pm}$ and Z^0. These particles are also of spin 1 but have masses of 80 and 91 GeV respectively. Weak interactions are responsible for radioactive beta decay of nuclei.
- The *gravitational* interactions take place between all forms of matter or radiation. They are mediated by the exchange of *gravitons*, of zero mass but spin 2.

For orientation on the magnitudes involved, the relative strengths of the different forces between two protons when just in contact are approximately

$$\begin{array}{cccc} \text{Strong} & \text{Electromagnetic} & \text{Weak} & \text{Gravitational} \\ 1 & 10^{-2} & 10^{-7} & 10^{-39} \end{array} \tag{1.4}$$

The weakness of gravity, compared with electromagnetism, is of course known to everyone from earliest childhood. As you fall down, you accelerate under the gentle force of gravity, but only on hitting the ground do you realise the enormously larger electromagnetic forces between molecules. Despite such differences, numerous attempts have been made over the years to find a unified theory, a so-called theory of everything. It turns out that the electromagnetic and weak interactions are in fact different aspects of a single *electroweak* interaction, as described below. The possibility of grander unification schemes and the reasons for them are discussed in Chapter 4.

Figure 1.3 shows diagrams depicting the above exchange processes, and Table 1.5 lists some of the properties of the interactions. In these diagrams (called in a more sophisticated form, Feynman diagrams after their inventor) solid lines entering or leaving the boundaries represent real particles—usually, quarks or leptons—with time flowing from left to right. The arrows along these lines indicate the direction of fermion number flow (see Fig. 1.9). An arrow indicating an electron flowing backwards in time is equivalent to a positron (i.e. antifermion) moving forwards: the convention is to use such time-reversed arrows for antiparticles.

Wavy, curly, or broken lines run between the vertices where the exchange interactions take place, and they represent the mediating bosons, which are *virtual* particles, that is, they carry energy and momentum such that the mass does not correspond to that of the free particle. To understand this, consider, for example, an electron of total energy E, momentum p, and mass m being scattered as another electron absorbs the exchanged photon. The relativistic

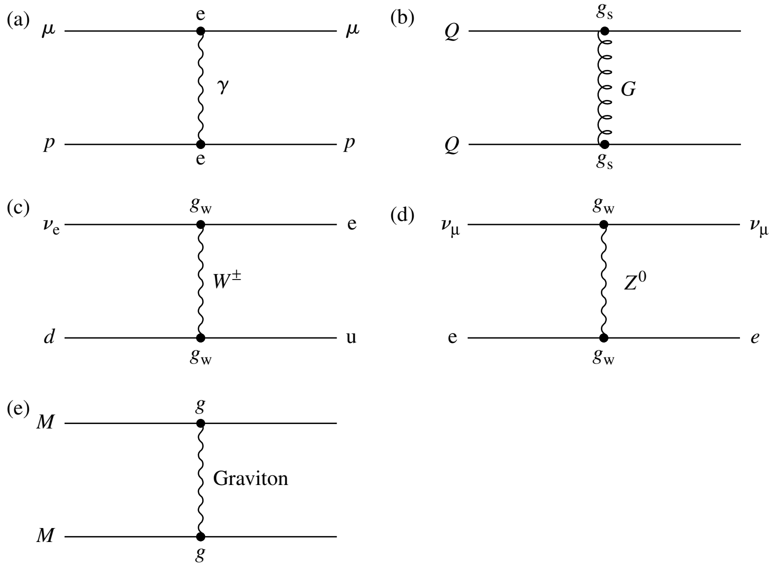


Fig. 1.3 Diagrams representing examples of single quantum-exchange processes in electromagnetic, strong, weak, and gravitational interactions. (a) The electromagnetic interaction between a muon μ and proton p, *via* photon (γ) exchange with coupling e. (b) The strong interaction between quarks Q *via* gluon (G) exchange with coupling g_s. (c) The weak interaction involving charged W boson exchange, transforming an electron-neutrino ν_e to an electron e, and a neutron (quark composition *ddu*) to a proton (*duu*). (d) The weak interaction involving neutral Z boson exchange, showing a muon–neutrino ν_μ scattering from an electron, e. In both (c) and (d), the couplings have been denoted g_W, but there are different numerical coefficients (of order unity) associated with the W and Z exchanges, as described in Chapter 3. (e) Gravitational interaction between two masses M, mediated by graviton (g) exchange. For macroscopic masses, multiple graviton exchanges will be involved.

relation between E, p, and m is given in (1.1), which in units $c = 1$ is

$$E^2 - p^2 = m^2$$

If the electron emits a photon of energy ΔE and momentum Δp then

$$E\,\Delta E - p\,\Delta p = 0$$

so that the mass of the exchanged photon is

$$\Delta m^2 = \Delta E^2 - \Delta p^2 = -\frac{m^2 \Delta p^2}{E^2} < 0 \tag{1.5}$$

Thus, as we know from common sense, a free electron cannot spontaneously emit a real photon, and the exchanged photon mass is imaginary—hence the term virtual. The energy ΔE has been 'borrowed' by the photon, and this is permitted for a time Δt limited by the Heisenberg uncertainty principle of quantum mechanics: $\Delta E\,\Delta t \sim \hbar$. However, if the virtual photon is absorbed by the second electron within the time Δt, energy and momentum balance can be satisfied. The quantity Δm^2 in (1.5) is defined in the rest-frame of the exchanged particle and is therefore a relativistically invariant quantity. As described in Chapter 2, it is usually called q^2, the square of the four-momentum transferred between the electrons.

Note that, if ΔE is large, Δt is correspondingly small and the range of the interaction $\Delta r \approx c\Delta t$ is correspondingly short. In 1935 Yukawa showed that the interaction potential $V(r)$ due to a spinless exchange boson of mass M had the form (see Appendix B)

$$V(r) \propto \left(\frac{1}{r}\right) \quad \exp\left(\frac{-r}{r_0}\right) \tag{1.6a}$$

where $r_0 = \hbar/Mc$, the Compton wavelength of the boson, is the effective range of the interaction. This result also follows from the uncertainty principle. A boson of mass M can exist as a virtual particle for a time $\Delta t \sim \hbar/Mc^2$, during which it can travel at most a distance $c\,\Delta t \sim \hbar/Mc$. The W and Z bosons have large masses and so the range r_0 is very short (of order 0.0025 fm). This is the reason that weak interactions are so much feebler than electromagnetic. The free photon associated with electromagnetism has rest-mass $M = 0$ and $r_0 = \infty$ so that the value of ΔE of the virtual photon can be arbitrarily small and the range of the interaction can therefore be arbitrarily large.

Instead of discussing the range of a static interaction, these features can be taken into account in a scattering process by defining a so-called *propagator*, measuring the amplitude for scattering with a momentum transfer q. Neglecting spin, this has the general form, following from the Yukawa potential (see Appendix B)

$$F(q^2) = \frac{1}{\left(-q^2 + M^2\right)} \tag{1.6b}$$

where $q^2 = \Delta E^2 - \Delta p^2$ is the (negative) four-momentum transfer squared in (1.5) and M is the (free particle) rest-mass of the exchanged boson. So for photons $M = 0$, for a weak boson $M = M_W$, and so on. The square of $F(q^2)$ enters into the cross-section describing the probability of the interaction, as discussed below. As an example, for Coulomb scattering $M = 0$ and the differential cross-section $\mathrm{d}\sigma/\mathrm{d}q^2$ varies as $1/q^4$, as given by the famous Rutherford scattering formula.

1.6 The boson couplings to fermions

1.6.1 Electromagnetic interactions

Apart from the effect of the boson propagator term, the strength of a particular interaction is determined by the coupling strength of the fermion (quark or lepton) to the mediating boson. For electromagnetic interactions, shown in Fig. 1.3(a), the coupling of the photon to the fermion is denoted by the electric charge $|e|$ (or fraction of it, in the case of a quark), and the product of the couplings of two fermions, each of charge $|e|$, to each other is

$$e^2 = 4\pi\alpha\hbar c,$$

more usually written as $4\pi\alpha$ in natural units with $\hbar = c = 1$. Here $\alpha \approx 1/137$ is the dimensionless fine structure constant. The cross-section or rate of a particular interaction is proportional to the square of the transition amplitude. This amplitude is proportional to the product of the vertex factors and the propagator term, that is, $\alpha/|q^2|$ for the electromagnetic interaction, corresponding to a factor $\alpha^2/|q^4|$ for the rate.

1.6.2 Strong colour interactions

For strong interactions, as in Fig. 1.3(b), the coupling of the quark to the gluon is denoted g_s with $g_s^2 = 4\pi\alpha_s$. Typically, α_s is a number of order unity. While in the electromagnetic interactions there are just two types of electric charge, denoted by the symbols + and −, as in Fig. 1.4(a), the interquark interactions involve six types of strong charge. This internal degree of freedom is called

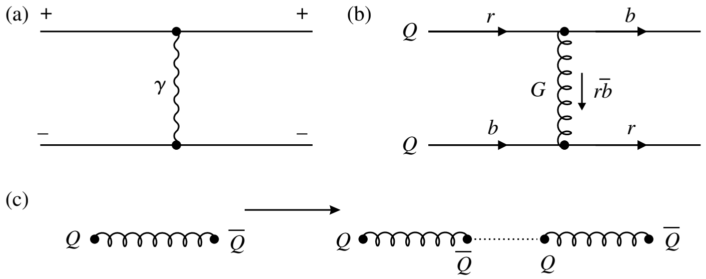


Fig. 1.4 The electromagnetic interaction in (a) involves two types of electric charge, + and − and is mediated by an uncharged photon. In (b) the strong interquark force involves six types of colour charge. The diagram depicts the interaction of a red quark with a blue quark via the exchange of a red–antiblue gluon. Diagram (c) depicts two quarks, connected by a gluon 'string', being pulled apart. Because of the confinement term in equation (1.7) the potential energy in the string grows linearly with the distance, and eventually it requires less energy to create a fresh quark–antiquark pair, involving two short strings, rather than one long one.

colour (nothing however to do with the optical spectrum). Quarks can carry one of three colours, say red, blue, or green, while antiquarks carry the anticolour. The quark combinations called hadrons have no colour. A baryon (proton or neutron) consists of one red, one blue, and one green quark, the combination being white (i.e. colourless). Similarly, a meson consisting of a quark of a particular colour and an antiquark of that anticolour, is also colourless.

Gluons, unlike photons, carry a colour charge, consisting of one colour and one anticolour. As an example, Fig. 1.4(b) shows a red quark r interacting with a blue quark b via the exchange of a $r\bar{b}$ gluon. The potential between two quarks due to the colour force is usually taken to be of the form

$$V(\text{colour}) = -\left(\frac{4}{3}\right)\frac{\alpha_s}{r} + kr \tag{1.7}$$

where r is the interquark separation, to be compared with the Coulomb potential between two unit charges of

$$V(\text{Coloumb}) = -\frac{\alpha}{r} \tag{1.8}$$

The factor 4/3 in (1.7) is a colour factor. Basically this comes about because there are eight possible colour–anticolour combinations of gluon ($3^2 = 9$, minus 1 which is a colourless singlet combination) to be divided between the six colours and anti-colours of quark and antiquark. Both potentials have a $1/r$ dependence at small distances, corresponding to the fact that both photons and gluons are massless. However, at larger distances the second term in (1.7) is dominant and is responsible for *quark confinement*. The value of k is about 0.85 GeV fm^{-1}.

Example 1.1 *Calculate the force in tonnes weight between a pair of quarks separated by a few fm.*

In equation (1.7) the attractive force at large r is $\mathrm{d}V/\mathrm{d}r = k =$ 0.85 GeV fm^{-1} or 1.36×10^5 J m^{-1}. Inserting the acceleration due to gravity, $g = 9.81$ m s^{-2}, a mass of 1000 kg exerts a force of 1 tonne-weight $= 9.8 \times 10^3$ J m^{-1}. Dividing, one finds $k = 13.9$ tonnes weight—a great deal for those tiny quarks, each weighing less than 10^{-24} gm.

Because gluons carry a colour charge (unlike photons which are uncharged) there is a strong gluon–gluon interaction. Thus the 'lines of colour force' between a pair of quarks, analogous to the lines of electric field between a pair of charges, are pulled out into a tube or string. In Fig. 1.4(c) such a gluon 'string' is depicted connecting a quark–antiquark pair. If one tries to pull apart the two quarks, the energy required to do so grows linearly with the string length as in (1.7), and eventually it requires less energy to produce another quark–antiquark pair, thus involving two short strings instead of one long one. Thus even the most violent efforts to separate quarks just result in production of lots of quark–antiquark pairs (mesons).

We may note at this point a peculiar property of the quark and lepton quantum numbers. Each of the three 'families' consists of a doublet of quarks of charge $+2/3|e|$ and $-1/3|e|$ respectively, and a pair of leptons with charges $-1|e|$ and 0. If allowance is made for the colour degree of freedom, the total charge of the quarks is $3 \times (2/3 - 1/3)\ |e| = +1|e|$ per family, while for the leptons it is $(-1\ \ + 0)|e| = -1|e|$. This is true for each family, so the total electric charge of the fermions is zero. In fact it turns out that an important quantity is the square of the electric charge multiplied by a quantity called the axial–vector coupling, which comes in with the same value but opposite signs for the charged and neutral leptons, and for the charge $+2/3$ and charge $-1/3$ quarks. Then in units of $|e|^2$ the total amounts to $[(0)^2 - (-1)^2] = -1$ for the leptons, and $3 \times \left[(+2/3)^2 - (-1/3)^2\right] = +1$ for the quarks, which again adds to zero. This turns out to be a crucial property, in making the theory free of so-called 'triangle anomalies' which, if they were not cancelled out, would spoil the renormalizability of the theory, discussed in Chapter 3. In fact, the fundamental reason for the existence of three families which are, so to speak, 'carbon copies' of one another, is at present unknown.

The confining force associated with the interquark potential (1.7) has dramatic effects in the process of high-energy electron–positron annihilation to hadrons. A first stage of the process is annihilation to a quark–antiquark pair, which in a second and separate stage transforms into hadrons, $\mathrm{e^+e^-} \to \mathrm{Q\bar{Q}} \to$ hadrons. The transverse momentum of a hadron is of order 0.3 GeV/c, that is, $\hbar/a$ where $a \sim 1$ fm is the force range, while the typical longitudinal momentum of a hadron from a high-energy collision is much larger, hence the usual appearance of two oppositely directed 'jets' of secondary particles. Such 'jets' of hadrons are the nearest that one ever gets to 'seeing' an actual quark. It may be remarked here that the observed cross-section for this process, compared with that for $\mathrm{e^+e^-} \to \mu^+\mu^-$ *via* photon exchange at the same energy, gave the first convincing evidence for the colour degree of freedom (see Fig. 1.10). Both are reactions proceeding *via* photon exchange, with a rate proportional to the square of the electric charges of the particles involved. The observed two jet event rate was consistent with that expected, provided a factor 3 enhancement was included, to take account of the fact that the quark–antiquark pair could be emitted in three colours $\left(r\bar{r}\ or\ b\bar{b}\ or\ g\bar{g}\right)$.

Figure 1.5 shows an event containing three, rather than two, jets of particles. In this case one of the quarks has radiated a high-energy gluon at wide angle, $\mathrm{e^+e^-} \to Q + \bar{Q} + G$. The ratio of three-jet to two-jet events clearly gives a measurement of the strong coupling α_s.

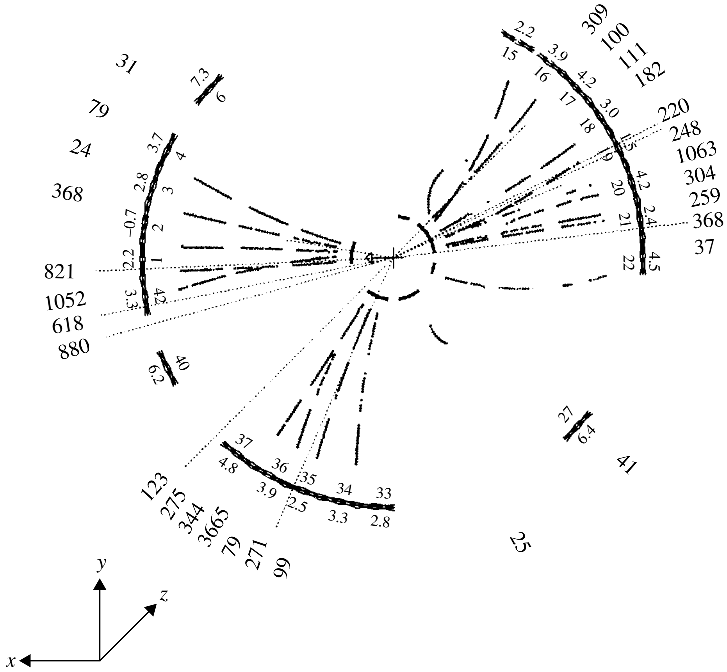


Fig. 1.5 Example of hadron production following e$^+$e$^-$ annihilation observed in the JADE detector at the PETRA collider at DESY, Hamburg. The total centre-of-momentum energy is 30 GeV. Trajectories of charged pions are shown as crosses, and of γ-rays from decay of neutral pions as dotted lines. The γ-rays are detected when they produce electron–photon showers in lead glass counters. Note the collimation of the hadrons into three distinct ‘jets’. (Courtesy DESY laboratory).

1.6.3 Weak Interactions

Figure 1.3(c) and (d) show the W and Z exchanges of the weak interactions, with couplings which we can generically denote by the symbol g_{w}. Thus the product of the propagator term and the coupling would give an amplitude in this case of

$$\underset{q^2\to 0}{Lt}\ \frac{g_{\mathrm{w}}^2}{\left(-q^2+Mw^2\right)} \equiv G_F \tag{1.9}$$

We have made the identification here with a point coupling G_F between the fermions involved, which Fermi had postulated in the earliest days of nuclear beta decay (1934). In those processes $|q^2| \ll M_W^2$ so that $G_F = g_{\mathrm{w}}^2/M_W^2$. The process of $W^{\pm}$ exchange in Fig. 1.3(c) involves a change in the charge of the lepton or quark, and is therefore sometimes referred to as a ‘charged current’ weak interaction, while the Z^0 exchange in Fig. 1.3(d) does not affect the charges of the particles and is termed a ‘neutral current’ weak interaction. While feeble compared with electromagnetic or strong interactions at GeV energies, the weak interactions play an extremely important role on the cosmic scale, precisely because of their weakness. As we shall see in later chapters, the way in which the universe evolved from very earliest times is greatly affected by the weak interactions of neutrinos, for example, in the development of large-scale galaxy clusters. At later times, the initial stages of stellar fusion processes are largely controlled by weak reactions which alone guarantee the long life of our own Sun.

1.6.4 Electroweak interactions

As stated above, the electromagnetic and weak interactions are *unified*. The *electroweak model* was developed in the 1960s, particularly by Glashow (1961), Salam (1967) and Weinberg (1967). Basically what this means is that the couplings of the W and Z bosons to the fermions are *the same as that of the photon*, that is, $g_w = e$. Here for simplicity we have omitted certain numerical factors of order unity. If one inserts this equality in (1.9) for the limit of low q^2, one obtains the expected large masses for the weak bosons:

$$M_{W,Z} \sim \frac{e}{\sqrt{G_F}} = \sqrt{\frac{4\pi\alpha}{G_F}} \sim 100 \text{ GeV} \tag{1.10}$$

where the value of $G_F = 1.17 \times 10^{-5}$ GeV^{-2} has been inserted from measured rate for muon decay (see also Table 1.5). So what (1.9) and (1.10) are telling us is that, although the photons and the weak bosons have the same couplings to leptons (to within a constant), the effective strength of the weak interaction is much less than that of the electromagnetic interaction because the larger mediating boson mass implies a much shorter range for the interaction. Because of this difference in boson masses, from zero for the photon to 80–90 GeV for the weak bosons, the electroweak symmetry is a broken symmetry. Figure 1.6 shows an example of the first observation of a W particle, in proton–antiproton collisions in 1983. Electroweak interactions are described in detail in Chapter 3.

1.6.5 Gravitational interactions

Finally, Fig. 1.3(e) depicts the gravitational interaction between two masses, via graviton exchange. The force between two equal point masses M is given by GM^2/r^2 where r is the separation and G is the Newtonian gravitational constant. Comparing with the electrostatic force between two charges $|e|$ of e^2/r^2, the quantity $GM^2/\hbar c$ is seen to be dimensionless. If we take as the unit of

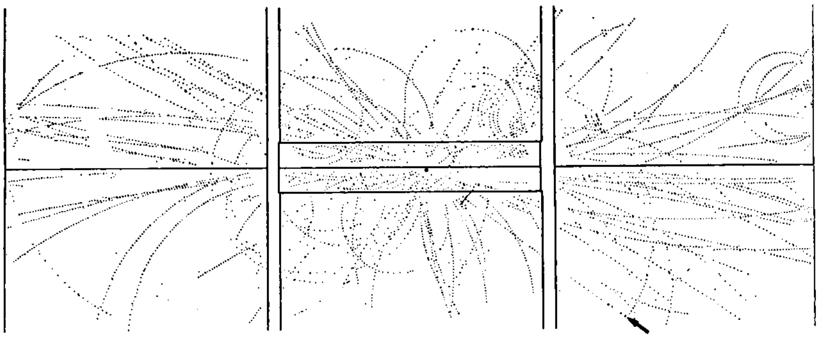


Fig. 1.6 One of the first examples of the production of a W boson at the CERN proton–antiproton collider in 1983. This reconstruction is of signals from drift chamber detectors surrounding the horizontal vacuum pipe. 270 GeV protons coming from the right collided with 270 GeV antiprotons from the left. Among the 66 tracks of secondary particles, one, shown by the arrow is an energetic (42 GeV) positron identified in a surrounding electromagnetic calorimeter. This positron has transverse momentum of 26 GeV/c, while the missing transverse momentum in the whole event is 24 GeV/c in the opposite sense, consistent with that of a neutrino, produced in the decay $W^{\pm} \rightarrow e^{+} + \nu_e$ (from Arnison *et al.* 1983).

mass $Mc^2 = 1$ GeV, then

$$\frac{GM^2}{4\pi\hbar c} = 5.3 \times 10^{-40} \tag{1.11a}$$

to be compared with

$$\frac{e^2}{4\pi\hbar c} = \frac{1}{137.036} \tag{1.11b}$$

Thus, for the energy or mass scales of GeV or TeV common in high-energy physics experiments at accelerators, the gravitational coupling is absolutely negligible. Of course, on a macroscopic scale, gravity is important and indeed dominant, because it is cumulative, since all particles with energy and momentum are attracted by their mutual gravitation. Thus the gravitational force on a charged particle on the Earth's surface is the sum of the attractive effects of all the matter in the Earth. Since the Earth is electrically neutral however, the enormously larger electrical force due to all the protons in the Earth is exactly cancelled by the opposing force due to the electrons.

However, even on sub-atomic scales the gravitational coupling can become strong for hypothetical elementary particles of mass equal to the *Planck mass*, defined as

$$M_{\mathrm{PL}} = \left(\frac{\hbar c}{\mathrm{G}}\right)^{1/2} = 1.2 \times 10^{19}\ \frac{\mathrm{GeV}}{\mathrm{c}^2} \tag{1.12a}$$

The *Planck length* is defined as

$$L_{\mathrm{PL}} = \frac{\hbar}{M_{\mathrm{PL}}c} = 1.6 \times 10^{-35}\ \mathrm{m} \tag{1.12b}$$

that is, the Compton wavelength of a particle of the Planck mass. Two pointlike particles each of the Planck mass and separated by the Planck length would therefore have a gravitational potential energy equal to their rest-masses, so quantum gravitational effects can become important at the Planck scale. To account for the very large value of the Planck mass, or the extreme weakness of gravity at normal energies, it has been proposed that there are extra dimensions beyond the familiar four of space/time, but these are 'curled up' to lengths of the order of the Planck length, so that they only become effective, and gravity becomes strong, at Planck energies.

We should emphasize here that, although we can draw a parallel between the inverse square law of force between point charges and point masses, there are quite fundamental differences between the two. First, due to the attractive force between two masses, the latter can acquire momentum and kinetic energy (at the cost of potential energy), which is equivalent to an increase in the effective mass through the Einstein relation $E = mc^2$, and thence in the gravitational force. For close enough encounters therefore, the force will increase faster than $1/r^2$. Indeed, one gets non-linear effects, which is one of the problems in formulating a quantum field theory of gravity. The effects of gravitational fields (including the non-linear behaviour) are enshrined in the Einstein field equations of general relativity, which interpret these effects in terms of the curvature of space caused by the presence of masses.

Second, it should be noted that in the above we have described gravity on a quantum-exchange basis, just like the other interactions. However, as discussed in Chapter 2, gravity is unique in that it cannot be simply treated as another field operating in normal 'flat' space. The Equivalence Principle equates the gravitational field with an accelerated reference frame, and so in a strong field, clocks run slow (or timescales are dilated), and lengths are contracted. Time and space, so to speak, get mixed up. Einstein (who of course pre-dated quantum mechanics) treated gravity as a geometrical property of space, and massive particles then altered the structure of space/time, and attracted one another because the space between them was 'curved' or 'warped'. Fine. But he then tried for years without success to bring other interactions, for example, electromagnetic interactions, into this picture. Unfortunately, attempts to unify the different interactions into a 'theory of everything', for example in the theory called 'supergravity', have not got very far.

While for the strong, electromagnetic and weak interactions, there is direct laboratory evidence for the existence of the mediating bosons—gluons, photons, and weak bosons respectively—so far the direct detection of gravitational waves (gravitons) has escaped us, although present experiments (2008) appear to be nearing the edge of success. Even the most violent events in the universe are expected to produce only incredibly small (10^{-22}) fractional deviations in detecting apparatus on Earth, caused by the compressing and extending effects of gravitational radiation. However, indirect evidence from the slow-down rate of binary pulsars discussed in Chapter 10, shows that gravitational radiation does indeed exist, and at exactly the rate predicted by general relativity. Gravitational radiation, or more precisely its imprint on the cosmic microwave background photons which emerged from the primordial 'soup' of matter and radiation when the universe was about 380,000 years old, appears to offer the only real possibility to 'see' further back to the very earliest, inflationary stages of our universe.

At this point we may note in passing that the gravitational, electromagnetic, and strong interactions can all give rise to (non-relativistic) *bound states*. A planetary system is an example of gravitational binding. Atoms and molecules are examples of binding due to electromagnetic interactions, while the strong interquark forces lead to three-quark (baryon) states as well as quark–antiquark bound states, for example the ϕ meson $s\bar{s}$ and the J/ψ meson $c\bar{c}$, appearing as resonances in electron–positron annihilation at the appropriate energy (see Fig. 1.10 below). Strong interactions are of course also responsible for the binding of atomic nuclei. The weak interactions do not lead to any bound states, because of the rapid decrease of the potential with distance as mentioned above.

Example 1.2 *Calculate at what separation r the weak potential between two electrons falls below their mutual gravitational potential.*

As indicated in (1.6), the Yukawa formula for the weak potential between two electrons, mediated by a weak boson of mass M and weak coupling g_{w} to fermions is

$$V_{\mathrm{wk}}(r) = \left(\frac{g_{\mathrm{w}}^2}{r}\right) \exp\left(\frac{-r}{r_0}\right) \quad \text{where } r_0 = \frac{\hbar}{Mc},$$

to be compared with the gravitational potential between two particles of mass m

$$V_{\text{grav}}(r) = \frac{\text{G}m^2}{r}$$

In the electroweak theory, $g_{\text{w}} \sim e$ so that $g_{\text{w}}^2/4\pi\hbar c \sim \alpha = 1/137$, while (see Table 1.5) $\text{G}m^2/4\pi\hbar c = 1.6 \times 10^{-46}$. Hence, inserting $M_{\text{w}} = 80$ GeV to obtain $r_0 = 2.46 \times 10^{-3}$ fm, one finds the two potentials are equal when $r \sim 100\, r_0 = 0.25$ fm.

To summarize this section, the characteristics of the fundamental interactions are listed in Table 1.5.

1.7 The quark–gluon plasma

As already stated, laboratory experiments indicate that quarks do not exist as free particles, but rather as three-quark and quark–antiquark bound states, called hadrons. There has long been speculation that this quark confinement mechanism is a low-energy phase and that at sufficiently high-energy densities, quarks and gluons might undergo a phase transition, to exist in the form of a plasma. An analogy can be made with a gas, in which at sufficiently high temperatures the atoms or molecules become ionised, and the gas transforms to a plasma of electrons and positive ions. If such a quark–gluon plasma is possible, the conditions of temperature and energy density in the very early stages (the first 25 μs) of the Big Bang would have certainly resulted in such a state existing, before the temperature fell as the expansion proceeded and the quark–gluon 'soup' froze out into hadrons.

Attempts have been made over the years to reproduce the quark–gluon plasma in the laboratory, by making head-on collisions of heavy nuclei (e.g. lead on lead) accelerated to relativistic energies. The critical quantity is the energy density of the nuclear matter during the very brief (10^{-23} s) period of the collision. For example, in lead–lead collisions at 0.16 GeV per nucleon in each of the colliding beams, a three-fold enhancement has been observed in the frequency of strange particles and antiparticles (from creation of $s\bar{s}$ pairs) as compared with proton–lead collisions at similar energy per nucleon.

Figure 1.7 shows the results of a compilation of data on the ratio of strange to non-strange particles as a function of centre-of-mass energy, from a review by Tannenbaum (2006). It is not clear at present whether the differences observed between nucleus–nucleus and nucleon–nucleon or electron–positron collisions are in fact due to such a phase transition. Obviously, present and future studies of such plasma effects in the laboratory (specifically at the RHIC heavy ion collider at Brookhaven National Laboratory, and the Large Hadron Collider at CERN), could be very important in shedding light on exactly how the early universe evolved.

1.8 The interaction cross section

The strength of the interaction between two particles, for example, in the two-body → two-body reaction

$$a + b \rightarrow c + d$$

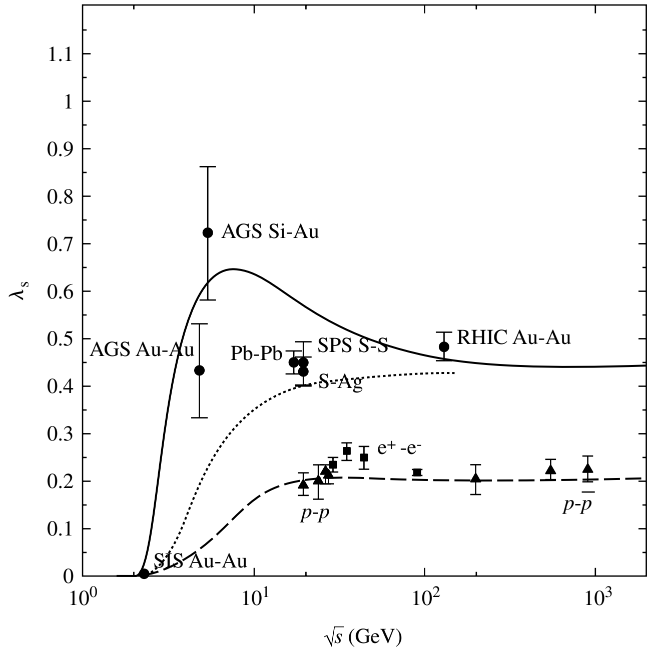


Fig. 1.7 The ratio of $s\bar{s}/(u\bar{u}+d\bar{d})$ quarks in pp, $p\bar{p}$, e^+e^-, and heavy ion collisions, as a function of centre-of-mass energy $\sqrt{s}$. The enhanced strange particle production in Au–Au, Pb–Pb, and Si–Au collisions is evident (from Tannenbaum 2006)).

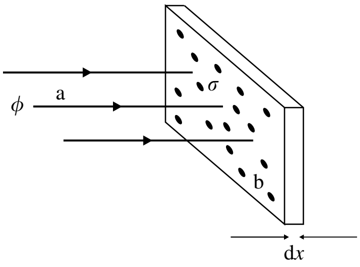


Fig. 1.8 Diagram indicating a beam of particles of type a incident on a target containing particles of type b.

is specified by the *interaction cross-section* σ defined as follows. Suppose the particles a are in a parallel beam, incident normally on a target of thickness dx containing n_b particles of type b per unit volume (see Fig. 1.8). If the density of incident particles is n_a per unit volume the flux—the number of particles per unit area per unit time—through the target will be

$$\phi_i = n_a\, v_i \tag{1.13}$$

where v_i is the relative velocity of beam and target. If each target particle has an effective cross-section of σ, then the fraction of the target area obscured, and the probability of collision, will be $\sigma n_b \mathrm{d}x$. The reaction rate per unit time per unit area of the target will then be $\phi_i \sigma n_b \mathrm{d}x$. Per target particle the reaction rate will be

$$W = \phi_i \sigma \tag{1.14}$$

so that the cross-section is equal to the reaction rate per target particle per unit incident flux. Cross-sections are measured in units called the *barn*. 1 barn = 1b = 10^{-28} m^2. This is roughly the geometric area of a nucleus of mass number $A = 100$. Appropriate units in particle physics are the millibarn (1 mb = 10^{-3} b), the microbarn (1 μb = 10^{-6} b), the nanobarn (1 nb = 10^{-9} b), and the picobarn (1 pb = 10^{-12} b).

The quantity W is given by an expression from (non-relativistic) perturbation theory, usually referred to as 'Fermi's Second Golden Rule' (derived in standard texts on atomic physics). It has the form

$$W = \left(\frac{2\pi}{\hbar}\right) |T_{\mathrm{if}}|^2 \rho_{\mathrm{f}} \tag{1.15}$$

where the transition amplitude or matrix element T_{if} between initial and final states is effectively an overlap integral over volume, $\int \psi_{\mathrm{f}}{}^* U \psi_{\mathrm{i}}\, \mathrm{d}V$, of the spatial parts ψ_i and ψ_f of the initial and final state wavefunctions, brought about by the interaction potential U. $\rho_\mathrm{f} = \mathrm{d}N/\mathrm{d}E_\mathrm{f}$ is the energy density of final states, which is the number of states in phase space available to the product particles per unit interval of the final state energy E_f. This may be found as follows.

Suppose a particle of arbitrary momentum p, described by a wavefunction ψ, is confined within a cubical box with perfectly reflecting sides of length L, with a volume $V = L^3$. What are the possible quantum states available to this particle? Since ψ must be single-valued, the number of de Broglie wavelengths $\lambda = h/p$ between the sides of the box must be an integer, say n. So for the x-component of momentum, $L/\lambda = Lp_x/h = n_x$, and the number of possible quantum states in the momentum interval $\mathrm{d}p_x$ will be $\mathrm{d}n_x = L\mathrm{d}p_x/h$. Similar expressions apply for the y- and z-components of momentum. So the overall number of possible states will be the product

$$\mathrm{d}N = \mathrm{d}n_x \mathrm{d}n_y \mathrm{d}n_z = \left(\frac{L}{h}\right)^3 \mathrm{d}p_x \mathrm{d}p_y \mathrm{d}p_z.$$

Since no direction in momentum space is preferred, the volume element $\mathrm{d}p_x\mathrm{d}p_y\mathrm{d}p_z = 4\pi p^2\mathrm{d}p$, and the number of states in the momentum interval $p \to p + \mathrm{d}p$ is

$$\mathrm{d}N = \left(\frac{V}{h^3}\right) 4\pi p^2 \mathrm{d}p.$$

The number of states in phase space available to a particle in the momentum interval $p \to p + \mathrm{d}p$, directed into solid angle $\mathrm{d}\Omega$ and enclosed in volume V is therefore

$$\mathrm{d}N = \frac{V p^2 \mathrm{d}p \mathrm{d}\Omega}{(2\pi\hbar)^3} \tag{1.16}$$

In the reaction $a + b \to c + d$, the final state wavefunction ψ_f will be the product wavefunction $\psi_c\psi_d$, so that to ensure that we end up with just one particle of each type when we integrate over volume in the transition matrix, a $V^{-1/2}$ normalization factor is needed for the wavefunction of each final state particle. When T_{if} is squared, the resulting $1/V$ factor cancels with the V factor in the phase space in (1.16), for each particle in the final state. Similarly, the normalization factors for the wavefunctions of the particles in the initial state cancel with the factors proportional to V for the incident flux and for the number of target particles. Hence the arbitrary normalization volume V cancels out, as indeed it must.

It is usual to express the cross-section in terms of quantities defined in the *centre-of-momentum system* (CMS) of the collision, that is, in a reference frame in which the vector sum of the three-momenta of the colliding particles is zero (see Chapter 2 for definitions and a discussion). This reference frame is chosen because it is relativistically invariant. Then from the above formulae one obtains (with $n_a = 1$ per normalization volume)

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \frac{W}{\phi_i} = \frac{W}{v_i} = \left(\frac{|T_{\mathrm{if}}|^2}{v_i}\right) p_f^2 \left(\frac{\mathrm{d}p_f}{\mathrm{d}E_f}\right) \left(\frac{1}{4\pi^2\hbar^4}\right) \tag{1.17}$$

where p_f is the numerical value of the oppositely directed momenta of c and d in the CMS and $E_f = E_c + E_d$ is the total energy in the CMS. Energy conservation gives

$$\sqrt{p_f^2 + m_c^2} + \sqrt{p_f^2 + m_d^2} = E_f$$

and thus

$$\frac{\mathrm{d}p_f}{\mathrm{d}E_f} = \frac{E_c E_d}{E_f p_f} = \frac{1}{v_f}$$

where v_f is the relative velocity of c and d. Then

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega}(a + b \rightarrow c + d) = \frac{1}{4\pi^2\hbar^4}\,|T_{\mathrm{i\,f}}|^2\,\frac{p_f^2}{v_{\mathrm{i}}\,v_{\mathrm{f}}} \tag{1.18}$$

We have so far neglected the spins s_a, s_b, s_c, and s_d of the particles involved. If a and b are unpolarized, that is, their spin substates are chosen at random, the number of possible substates for the final state particles is $g_{\mathrm{f}} = (2s_{\mathrm{c}} + 1)$ $(2s_{\mathrm{d}}+1)$ and the cross-section has to include the factor g_{f}. For the initial state the factor is $g_i = (2s_{\mathrm{a}} + 1)\ (2s_{\mathrm{b}} + 1)$. Since a given reaction has to proceed through a particular spin configuration, one must average the transition probability over all possible initial states, all equally probable, and sum over all final states. This implies that the cross-section has to be multiplied by the factor $g_{\mathrm{f}}/g_{\mathrm{i}}$.

The reduction in the number of incident particles $n = n_a$ after passage through a thickness dx of absorber in Fig. 1.8 is

$$\mathrm{d}n = -n\sigma\rho\ \mathrm{d}x$$

where $\rho = n_b$ is the density of target particles. Integrating we obtain

$$n(x) = n(0)\ \exp(-\sigma\rho x) \tag{1.19}$$

Thus the proportion of incident particles which survive without interaction falls to $1/e$ in a distance

$$\lambda = \frac{1}{(\sigma\rho)} \tag{1.20}$$

The quantity λ is called the *mean free path* for interaction. It is left as an exercise to show that, from the distribution (1.19) and the definition (1.20), λ is the mean path length between collisions.

1.9 Examples of elementary particle cross sections

The two-body to two-body reactions described above are important in discussing the role of elementary particle interactions in the early universe, and we give here some examples. A full evaluation of the cross-sections would in some cases involve rather lengthy Dirac algebra, so that all that we shall do here is to give approximate expressions which can be justified simply on dimensional grounds, and which will give orientation on the dependencies and magnitudes involved. For simplicity we also neglect considerations of spin at this stage. Spin and helicity factors are taken into account later in the text, when exact cross-sections are required.

As (1.18) indicates, the differential cross-section for an extreme relativistic two-body to two-body elastic collision has the form, in units $\hbar = c = 1$

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = |T_{\mathrm{if}}|^2 \, \frac{s}{64\pi^2} \qquad \big(\text{extreme relativistic two-body} \to \text{two-body}\big) \qquad (1.21)$$

where $s = E_f^2$ denotes the square of the CMS energy, and $E_f = 2p_f$, since the masses involved are small compared with the energies, and $v_i = v_f = 2$. In this case also, the four-momentum transfer squared is $|q^2| = 2p_f^2(1 - \cos\theta)$ where θ is the angle of emission of the secondary particles in the CMS, so that $\mathrm{d}q^2 = p_f^2 \mathrm{d}\Omega/\pi$ and (1.21) can also be written:

$$\frac{\mathrm{d}\sigma}{\mathrm{d}q^2} = \frac{|T_{\mathrm{if}}|^2}{16\pi} \qquad (1.22)$$

Examples of electromagnetic cross-sections are as follows:

(a) $\mathbf{e^-\mu^+ \to e^-\mu^+}$

This is an example of the Coulomb scattering between singly charged leptons, as shown in Fig. 1.9(a). The couplings and photon propagator term together give $|T_{\mathrm{if}}| = e^2/\left|q^2\right|$ so that

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} \sim \frac{\alpha^2 s}{q^4} \qquad (1.23)$$

where $\alpha = e^2/4\pi$. This is just the Rutherford formula for pointlike scattering, and if θ is the angular deflection of the incident particle, then $q = 2p \sin(\theta/2)$ and one obtains the famous cosec^4 $(\theta/2)$ dependence on scattering angle

(b) $\mathbf{e^+e^- \to \mu^+\mu^-}$

The diagram for this process, in Fig. 1.9(b), is just that in (a) rotated through 90° and with outgoing leptons replaced by incoming antileptons and vice-versa. These two diagrams are said to be 'crossed' diagrams. In this case $|q^2| = s$, the

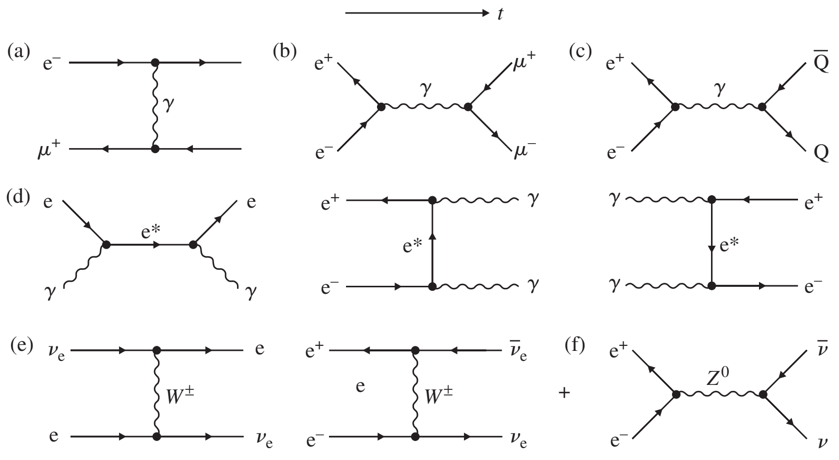


Fig. 1.9 Feynman diagrams for various elementary two-body to two-body reactions. In these diagrams. Time flows from left to right. The convention is that right-pointing arrows denote particles, while left-pointing arrows denote antiparticles. Diagrams (a) to (d) refer to electromagnetic interactions, and (e) and (f) to weak interactions. (a) $e^-\mu^+ \to e^-\mu^+$; (b) $e^+e^- \to \mu^+\mu^-$; (c) $e^+e^- \to Q\bar{Q} \to$ hadrons; (d) $e\gamma \to e\gamma$, $e^+e^- \to \gamma\gamma$, $\gamma\gamma \to e^+e^-$; (e) $\nu e \to \nu e$; (f) $e^+e^- \to \nu\bar{\nu}$.

square of the CMS energy, so that

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} \sim \frac{\alpha^2}{s} \tag{1.24}$$

In fact a full calculation gives for the total cross-section

$$\sigma = \frac{4\pi\alpha^2}{3s} \tag{1.25}$$

This is the cross-section based upon single photon exchange, as shown by the dashed line in Fig. 1.10. There will also be a contribution from Z^0 exchange, but because of the propagator term in (1.9), this is strongly suppressed at GeV energies. The above result is also expected on dimensional grounds. In units $\hbar = c = 1$, the cross-section has dimensions of GeV^{-2} and if the CMS energy dominates over the lepton masses involved, the $1/s$ dependence must follow. Since, from Table 1.1, 1 $\mathrm{GeV}^{-1} = 1.975 \times 10^{-16}$ m, the above cross-section is readily calculated to be 87/s nb (with s in GeV^2).

(c) $\mathbf{e^+e^- \rightarrow Q\bar{Q} \rightarrow}$ hadrons

The same formula (1.24) applies, again assuming the quarks have relativistic velocities, replacing the unit charges by the fractional quark charges at the right-hand vertex (Fig. 1.9(c)), and multiplying by a factor 3 for the number of quark colours. Of course one does not observe the actual quarks: they 'fragment' into hadrons by gluon exchanges as shown, but this is a relatively slow and independent second-stage process. Except near meson resonances, the cross-section is determined entirely by the elementary $\mathrm{e^+e^-} \rightarrow Q\bar{Q}$ process.

Figure 1.10 shows a plot of the cross-section as a function of energy. The peaks are due to the excitation of bound quark–antiquark states or resonances forming short-lived mesons: for example, the ρ and ω-mesons formed from u and d quarks and antiquarks; the ϕ formed from $s\bar{s}$; the J/ψ formed from $c\bar{c}$, the Υ from $b\bar{b}$, and finally the Z^0 resonance. Nevertheless, the general $1/s$ dependence of the cross-section, aside from these resonance effects, is quite clear.

Example 1.3 *Calculate the ratio R of the cross-section for* $\mathrm{e^+e^-} \rightarrow Q\bar{Q} \rightarrow$ hadrons, *to that for* $\mathrm{e^+e^-} \rightarrow \mu^+\mu^-$, *via photon exchange, as a function of CMS energy (see also Fig. 1.10).*

The quark–antiquark cross-section is proportional to the square of the quark charges and carries a factor 3 for colour. Thus for $u\bar{u}$, $d\bar{d}$, and $s\bar{s}$ quarks, the factor is

$$R = 3 \times \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right] = 2$$

and this is shown as the first 'step' above the pointlike cross-section for a muon pair in Fig. 1.10. Above the charmed ($c\bar{c}$) threshold appears a second step with $R = 11/3$, and above the $\mathrm{b\bar{b}}$ threshold at CMS energy of 10 GeV, $R = 11/3$.

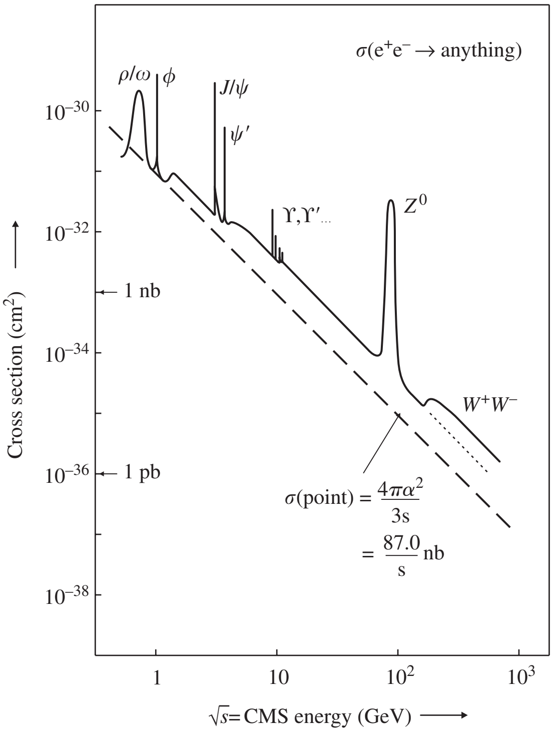


Fig. 1.10 The cross-section for the reaction $e^+e^- \to$ anything, as a function of CMS energy. The prominent peaks are due to various boson resonances, as described in the text. The overall $1/s$ dependence, typical of a pointlike scattering process, is clear. The magnitude of the cross-section for $e^+e^- \to$ anything, compared with that for muon pair production, shown as a dashed line, provides evidence for a factor 3 for the hadronic cross-section due to the colour degree of freedom of the quarks. The data are taken from measurements at various electron–positron colliders, the largest of which was the LEP collider at CERN, Geneva, in which 100 GeV electrons collided head-on with 100 GeV positrons.

Example 1.4 *Estimate an approximate value for the cross-section for the production of the top quark, of mass $m_t = 175$ GeV/c^2, in proton–antiproton collisions at centre-of-mass energies large compared with the quark mass.*

The top quark was discovered in 1995 at the Fermilab proton–antiproton collider, with a collision energy of 1.8 TeV in the centre-of-mass system (CMS). The principal process is that of production of a $t\bar{t}$ pair in the collision of a u or d quark from the proton with an antiquark from the antiproton, *via* gluon exchange, that is, $u + \bar{u} \to t + \bar{t}$, as in Fig. 1.9(c) but with incident e^+e^- replaced by $u\bar{u}$ (or $d\bar{d}$) and gluon exchange with coupling α_s replacing photon exchange with coupling α. Then from (1.24) we expect the cross-section to be $\sigma \sim F\alpha_s^2/s$, where $s \sim (2m_t)^2$, that is, assuming an incident $Q\bar{Q}$ centre-of-mass energy just above threshold, with F representing the probability that the colliding quarks are above threshold. If we set $\alpha_s = 0.1$ as a typical value, $m_t = 175$ GeV, then $\sigma \sim 30F$ pb. The value of F depends on the momentum distribution of quarks in the nucleon. A detailed calculation ends up with a cross-section of 7 pb, in agreement with observation.

Since the total proton–antiproton collision cross-section at these energies is 80 mb, this meant that the top quark was produced in only about 1 in 10^{10} collisions. Despite this huge background, top quarks could be detected because of the very distinguishing features of their decays, for example $t \to W^+ + b$ and $\bar{t} \to W^- + \bar{b}$. The W-bosons decay to give muons at wide angle and neutrinos, which manifest themselves as 'missing' energy and momentum, while the b quarks produce hadronic jets, slightly displaced from the main vertex because of the finite lifetime of the *B*-mesons. The signal is so characteristic and specific that the background could be reduced below the 10% level. One of the detectors used in discovering the top quark is shown in Fig. 1.11.

(d) $\mathbf{e^+e^- \to \gamma\gamma,\ \gamma\gamma \to e^+e^-,\ \gamma e \to \gamma e}$

These are important processes in astrophysics. They can all be represented by the same Feynman diagram, as in Fig. 1.9(d). Because in this case a virtual electron, rather than photon, operates between the vertices, the formula (1.24) is modified by a logarithmic term. Of course, although represented by the same diagram, the three processes have different dynamics (thresholds). In the limit of high energy, that is, for CMS energy squared $s \gg m^2$ where m is the electron mass, the cross-sections have the asymptotic forms:

$$\sigma\left(e^+e^- \to \gamma\gamma\right) = \left(\frac{2\pi\alpha^2}{s}\right)\left[\ln\left(\frac{s}{m^2}\right) - 1\right] \tag{1.26a}$$

$$\sigma\left(\gamma\gamma \to e^+e^-\right) = \left(\frac{4\pi\alpha^2}{s}\right)\left[\ln\left(\frac{s}{m^2}\right) - 1\right] \tag{1.26b}$$

$$\sigma\left(\gamma e \to \gamma e\right) = \left(\frac{2\pi\alpha^2}{s}\right)\left[\ln\left(\frac{s}{m^2}\right) + \frac{1}{2}\right] \tag{1.26c}$$

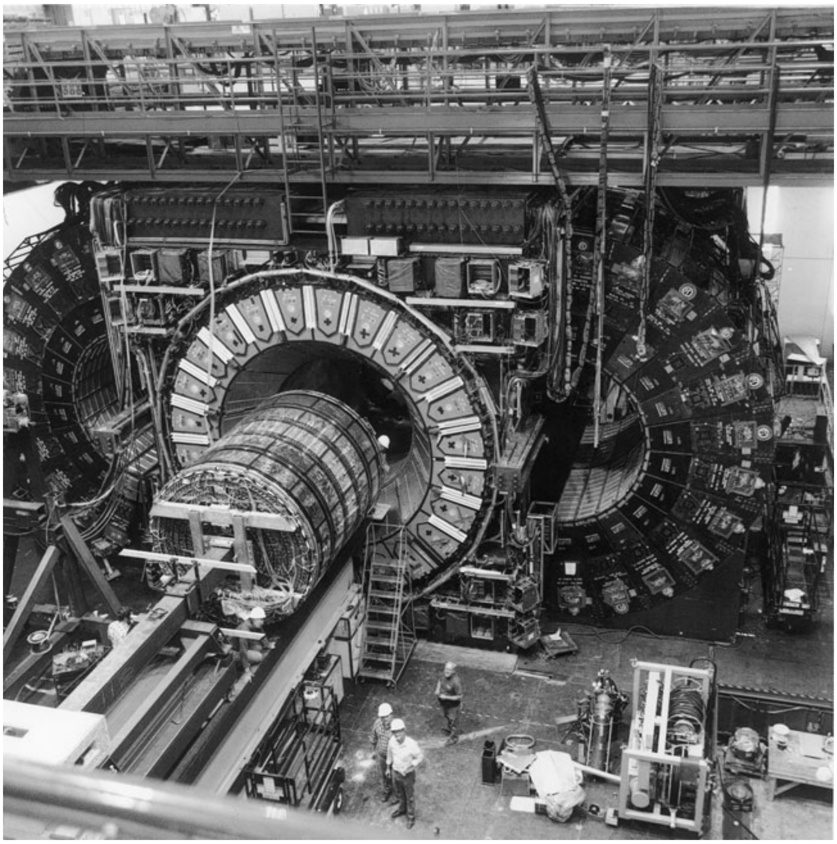


Fig. 1.11 Photograph of the CDF detector employed in the discovery of the top quark in 1995, in collisions of 0.9 TeV protons with 0.9 TeV antiprotons at the proton–antiproton collider at Fermilab, near Chicago. In the centre of the picture is the central tracking detector, which records the trajectories of individual secondary particles in drift chambers, and measures their momenta from track curvature in the applied magnetic field. Inside this are precision solid-state (silicon strip) detectors which can record tracks and secondary vertices very close to the main interaction vertex. Surrounding the central tracker are calorimeters which measure the total energy in charged and neutral particles. They are built in two arches which are withdrawn in the photograph. Outside the calorimeter modules and magnet yoke are further chambers to record penetrating muons from the annihilation reactions. (Courtesy Fermilab Visual Media Services).

The fact that the cross-section (1.26a) is half that of (1.26b) arises because in the first process there are two indistinguishable particles in the final state, so that the phase–space volume is halved. At collision energies large compared with the W mass, a formula similar to (1.26a) would apply to the reaction $e^+e^- \to W^+W^-$. The last process (1.26c) is known as Compton scattering. The same formula applies when the roles of incident and target particle are reversed, that is, γ-rays are accelerated to higher energies following collision with incident electrons. This inverse Compton effect is believed to be important in the production of high-energy γ-rays from point stellar sources.

At the other extreme of low CMS energy, the electron mass will dominate the energy scale and s is replaced by m^2, so that $\sigma \sim \alpha^2/m^2$. In fact the classical Thomson cross-section, applying for Compton scattering as $E_\gamma \to 0$, has the value

$$\sigma\,(\gamma e \to \gamma e)_{\text{Thomson}} = \frac{8\pi\alpha^2}{3m^2} = 0.666 \text{ barns} \tag{1.26d}$$

For the process $\gamma\gamma \to e^+e^-$, the threshold energy is $s_{\text{th}} = 4\text{m}^2$ and $\sigma \sim \beta\alpha^2/s$, where $\beta = \left(1 - 4m^2/s\right)^{1/2}$ is the CMS velocity of the electron or positron. Thus the cross-section at first increases with energy, and reaches a maximum of about $0.25\sigma_{\text{Thomson}}$ at $s \sim 8m^2$, before falling off at higher s values.

In summary, the above processes involve massless or almost massless photon or electron propagators and their effect is to introduce a $1/s$ dependence to the cross-sections. We now discuss weak processes which involve the massive W and Z propagators.

(e) $\mathbf{\nu_e e \to \nu_e e;\ e^+e^- \to \nu_e\bar{\nu}_e}$

Figure 1.9(e) shows the diagram for neutrino–electron scattering via $W-$exchange. There is also a contribution from Z exchange, but we consider here only the former. The propagator gives a term $1/\left(\left|q^2\right| + M_W^2\right)$ in $|T_{\text{if}}|$. Because of the large value of the W mass (80 GeV), $|q^2| \ll M_W^2$ at normal (<1 TeV) neutrino energies and therefore

$$\sigma \sim \frac{g_w^4 s}{M_W^4} \sim \text{G}_\text{F}^2 s \tag{1.27a}$$

where G_F is the Fermi constant defined in (1.9). An exact calculation gives

$$\sigma\,(\nu_e e \to \nu_e e) = \frac{\text{G}_\text{F}^2 s}{\pi} \tag{1.27b}$$

again assuming that lepton masses can be neglected in comparison with the collision energy. For the 'crossed' reaction $e^+e^- \to \nu_e\bar{\nu}_e$ shown in Fig. 1.9(f), the cross-section is (see Section 3.6):

$$\sigma\left(e^+e^- \to \nu_e\bar{\nu}_e\right) = \frac{\text{G}_\text{F}^2 s}{6\pi} \tag{1.28}$$

In addition to $W^\pm$ exchange, this reaction can also proceed through Z^0 exchange, and indeed this is the only possibility for the reactions $e^+e^- \to \nu_\mu\bar{\nu}_\mu$

or $\nu_\tau \bar{\nu}_\tau$. These processes are also discussed in Section 3.6. They are of astrophysical significance, both in the very early stages of the universe, and in the later supernova stages of giant stars.

1.10 Decays and resonances

As indicated in (1.15), an unstable state has a decay rate W, usually quoted as a *width* Γ in energy units, which corresponds to the fact that a non-stationary state with a finite lifetime must have a spread in energy, in accord with the uncertainty relation, that is $\Gamma = \hbar W = \hbar/\tau$ where $\tau = 1/W$ is the mean lifetime of the state.

As an example of a decaying state, let us consider muon decay, $\mu^+ \rightarrow e^+ + \nu_e + \bar{\nu}_\mu$. Clearly the transition amplitude for this weak decay $|T_{\mathrm{if}}| \propto G_{\mathrm{F}}$, the Fermi coupling constant, which has dimensions (energy)$^{-2}$—see Table 1.5. Hence the square of the transition amplitude has dimensions (energy)$^{-4}$ while the decay rate or width has dimensions of energy. Thus by dimensional arguments the phase–space factor in (1.15) must vary as (energy)^5, and since the largest energy or mass involved is $m_\mu c^2$ it follows that the muon decay rate $\Gamma \sim G_{\mathrm{F}}^2 m_\mu^5$. In fact a full (and quite lengthy) calculation gives

$$\Gamma\left(\mu^+ \rightarrow e^+ \nu_e \bar{\nu}_\mu\right) = \frac{G_{\mathrm{F}}^2 m_\mu^5}{192\pi^3} \tag{1.29}$$

It is left as an exercise to verify the value of the Fermi constant in Table 1.5 from the measured lifetime $\tau_\mu = \hbar/\Gamma = 2.197$ μs and mass $m_\mu c^2 = 105.66$ MeV (one should remark that the result will not be quite exact, because there are radiative corrections at the per cent level).

One can understand that a state with a measurable lifetime usually has an unmeasurably small width, while one with a broad and measurable width, for example, one decaying through the strong interactions, usually has an unmeasurably short lifetime, and is referred to as a *resonance*. Such resonances can readily be formed in collisions between the particles into which they decay. The exponential nature of the time distribution of decays determines the form of the line shape of the resonance.

Denoting the central frequency of the resonant state by ω_{R}, the wavefunction describing this state can be written

$$\psi(t) = \psi(0)\ \exp(-i\omega_{\mathrm{R}} t)\ \exp\left(-\frac{t}{2\tau}\right) = \psi(0)\ \exp\left\{\frac{-t\left(iE_{\mathrm{R}} + \left(\Gamma/2\right)\right)}{\hbar}\right\} \tag{1.30}$$

where the central energy is $E_{\mathrm{R}} = \hbar\omega_{\mathrm{R}}$ and the width $\Gamma = \hbar/\tau$. The intensity $I(t) = \psi^*(t)\ \psi(t)$ obeys the usual radioactive decay law

$$\frac{I(t)}{I(0)} = \exp\left(\frac{-\Gamma t}{\hbar}\right) \tag{1.31}$$

The energy dependence of the cross-section for forming the resonance is the Fourier transform of the time pulse, in the same way that, in wave optics, the angular distribution of the beam diffracted by a slit system is the Fourier transform of the slit profile. The Fourier transform of (1.30) is

$$g(\omega) = \int \psi(t)\ \exp(i\omega t)\ \mathrm{d}t$$

With $E = \hbar\omega$ the amplitude as a function of E is then (in units $\hbar = c = 1$)

$$A(E) = \psi(0) \int \exp\left\{-t\left[\left(\frac{\Gamma}{2}\right) + i\,(E_{\mathrm{R}} - E)\right]\right\} \mathrm{d}t = \frac{K}{\left\{(E - E_{\mathrm{R}}) - \left(i\Gamma/2\right)\right\}} \tag{1.32}$$

where K is some constant. The cross-section $\sigma(E)$, measuring the probability of two particles a and b forming a resonant state c will be proportional to A^*A, that is,

$$\sigma(E) = \sigma_{\max} \frac{\Gamma^2/4}{\left[(E - E_{\mathrm{R}})^2 + \left(\Gamma^2/4\right)\right]} \tag{1.33}$$

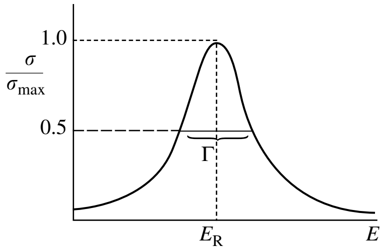


Fig. 1.12 The Breit–Wigner resonance curve.

which is called the *Breit–Wigner resonance formula*. The shape of the resonance is shown in Fig. 1.12, from which we note that the cross-section falls to half its peak value for $E - E_{\mathrm{R}} = \pm\Gamma/2$. The value of the peak cross-section in (1.33) can be evaluated as follows.

An incident particle of momentum p will be described by a plane wave, which can be decomposed into a superposition of spherical waves of different angular momentum l with respect to the scattering centre , where $l\hbar = pb$ and b is the 'impact parameter'. Particles of angular momentum in the interval $l \to l + 1$ therefore impinge on an annular ring of cross-sectional area

$$\sigma = \pi\left(b^2_{(l+1)} - b^2_l\right) = \pi\bar{\lambda}^2\,(2l + 1) \tag{1.34}$$

where $\bar{\lambda} = \hbar/p$. If the scattering centre is totally absorbing, σ in (1.34) will be the reaction or absorption cross-section. More generally, we can write the radial dependence of the outgoing amplitude (for the lth partial wave) in the form

$$r\ \psi(r) = \exp[ikr]$$

so that the total flux $4\pi r^2|\psi(r)|^2$ through radius r is independent of r. This is for the case of no scattering centre, while if the scattering centre is present

$$r\ \psi(r) = \eta \exp[i(kr + \delta)]$$

Here $0 < \eta < 1$ and δ is a phase shift. By conservation of probability, the reaction cross-section σ_r will be given by the difference of intensities with and without the scattering,

$$\sigma_r = \sigma\left(1 - \eta^2\right).$$

The scattered amplitude will clearly be

$$A = \exp[ikr] - \eta \exp[i(kr + \delta)]$$

giving for the scattered intensity

$$A^*A = 1 + \eta^2 - 2\eta \cos\,\delta.$$

Thus for $\eta = 0$ (total absorption) both elastic and reaction cross-sections are equal to σ in (1.34). The elastic cross-section in this case corresponds to the elastically diffracted beam from the absorbing obstacle. The other extreme case is that of pure scattering ($\eta = 1$) without absorption but just a shift in phase. Then

$$\sigma_{\text{el}} = 4\sigma \sin^2\left(\frac{\delta}{2}\right)$$

The maximum effect is for a phase shift of π radians, leading to a scattering amplitude equal to twice that for total absorption, or a cross-section

$$\sigma_{\text{el}}(\max) = 4\pi\lambda\!\!\!^{-2}(2l+1) \tag{1.35}$$

So far, we have omitted the effects of particle spin. The appropriate spin multiplicity factors were given in Section 1.8 above. Putting all these things together, the complete Breit–Wigner formula becomes

$$\sigma(E) = \frac{4\pi\lambda\!\!\!^{-2}\,(2J+1)\left(\Gamma^2/4\right)}{(2s_a+1)\,(2s_b+1)\left[(E-E_{\text{R}})^2 + \left(\Gamma^2/4\right)\right]} \tag{1.36}$$

where s_a and s_b are the spins of the incident particles and J is the spin of the resonant state (all in units of $\hbar$). Usually, the resonance from the reaction $a + b \to c$ can decay in a number of modes, each one with a *partial width* Γ_i for the *i*th mode, so that the fractional probability of decaying through that mode is Γ_i/Γ. In general, the resonance is formed through channel *i* and decays through channel *j*, and the cross-section is then given by multiplying (1.36) by the ratio $\Gamma_i\Gamma_j/\Gamma^2$.

1.11 Examples of resonances

We cite here some examples of resonances, of importance both in particle physics and in astrophysics. Figure 1.13 shows the cross-section for the process $\text{e}^+\text{e}^- \to$ anything, in the neighbourhood of the Z^0 resonance (the Z^0 being the mediator of the neutral current weak interaction). The central mass $E_{\text{R}} =$ 91 GeV and the total width Γ = 2.5 GeV. This resonance has many possible decay modes; into hadrons *via* pairs of *u, d, s, c*, or *b* quarks and antiquarks, into pairs of charged leptons e^+e^-, $\mu^+\mu^-$, or $\tau^+\tau^-$, or into neutrino pairs $v_e\bar{v}_e$, $v_\mu\bar{v}_\mu$, or $v_\tau\bar{v}_\tau$. At the time that this resonance was first investigated, there was some question about the total number of families of quarks and leptons (and the top quark had not yet been discovered). Could there be more than three types or flavours of neutrino? The curves in Fig. 1.13 show the effect on the width of assuming 2, 3, or 4 flavours of neutrino, based on the couplings of the Z to quarks and leptons as prescribed by the Standard Model. Clearly the observed width bears out the Standard Model assumption of three families. The number of flavours of neutrino, as discussed in Chapter 6, has an effect on the primordial helium/hydrogen ratio and thence on the subsequent evolution of the stars.

Example 1.5 *Calculate the peak cross-section for the production of the* Z^0 *resonance in the reaction* $\text{e}^+ + \text{e}^- \to Z^0$, *where the partial width is given by* $\Gamma_{\text{ee}}/\Gamma_{\text{total}} = 0.033$. *Compare the answer with the result in Fig. 1.13.*

Inserting $\lambda\!\!\!^- = \hbar c/pc$ where the CMS momentum $pc = M_Z/2 =$ 45.5 GeV, and with $J = 1$, $s = 1/2$, and $\Gamma_{ee}/\Gamma_{\text{tot}} = 0.033$, one obtains

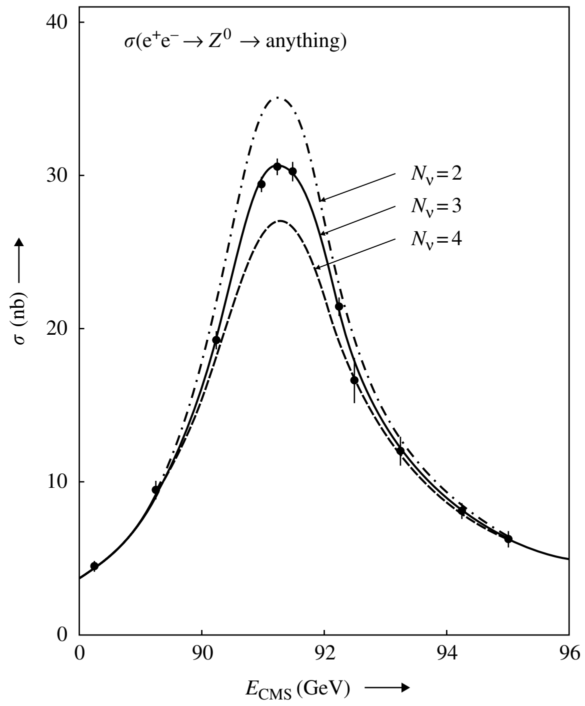


Fig. 1.13 The electron–positron annihilation cross-section as a function of energy near the Z^0 resonance. The observed values are averages from four experiments at the LEP electron–positron collider at CERN. The three curves are the Standard Model predictions for 2, 3, or 4 flavours of neutrino. In this case the Breit–Wigner curve is asymmetric, because the nominal beam energy is modified by synchrotron radiation losses.

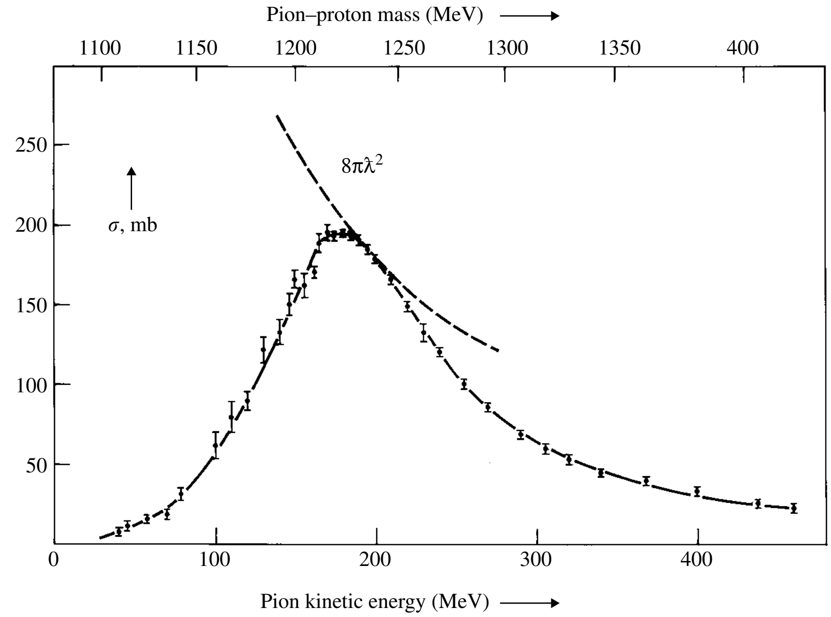


Fig. 1.14 The pion-proton resonance Δ (1232) first observed by Anderson *et al.* in 1952.

from (1.32), $\sigma(\text{peak}) = \left[12\pi/M_Z^2\right] \times 0.033(\hbar c)^2 = 58\,\text{nb}$. The actual cross-section (Fig. 1.13) is about half of this. The difference is due to radiative corrections to the electron and positron in the initial state, which smear the energy distribution and depress the peak cross-section.

Figure 1.14 shows the first resonance ever to be discovered in high-energy physics, namely the $\Delta(1232)$ pion-proton resonance observed in 1952. It has central mass 1232 MeV/c^2 and width $\Gamma = 120$ MeV. This observation was followed by that of many other meson–meson and meson–baryon resonances in the 1950s and 1960s. These resonant states were important in providing the essential clues which led to the development of the quark model by Gell-Mann and by Zweig in 1964. The Δ resonance is also of present astrophysical significance in connection with the very highest-energy cosmic rays, since it can be excited in collisions of protons above 10^{19} eV energy with the cosmic microwave background, with quantum energy of order of 0.25 meV (milli-electron volt). Known as the GZK effect after Greisen, Zatsepin, and Kuzmin who suggested it, the resonance indeed leads to a cut off in the cosmic ray spectrum above 10^{19} eV (see Section 9.12).

From the viewpoint of the human race, possibly the most important resonance is that of the 0^+ excited state of the ^{12}C nucleus at an excitation energy of 7.654 MeV. The width is about 10 eV only. The production of carbon in helium-burning red giant stars, discussed in Chapter 10, is achieved through the so-called triple alpha process, $3\alpha \rightarrow {}^{12}\mathrm{C}$. First, two alpha particles combine to form ^{8}Be in its ground state, which is unstable with a lifetime of only 10^{-16} sec. It may nevertheless capture a third alpha particle to form carbon, which, however, usually decays back into beryllium plus an alpha particle, but can with a small (10^{-3}) probability decay by gamma emission to the ground state of carbon. The rate of ^{12}C production depends crucially on the existence of this resonance level, occurring just 400 keV above the threshold energy, to enhance the triple alpha cross-section. Indeed, Hoyle had predicted the need for such a resonance and its properties in 1953, before it was finally found in laboratory experiments. Without the existence of this resonance, it is almost certain that carbon-based biological evolution in the universe could never have taken place.

1.12 New particles

The Standard Model of the fundamental quarks and leptons and their interactions described here and in Chapter 3 is able to account for practically all laboratory experiments at accelerators to date, and describes with great accuracy the physics of the fundamental particles, at least in our particular corner of the universe. However, it apparently does not describe the building blocks of the universe on large scales. Indeed on such scales it can account for only about 4% of the total energy density! As described in Chapter 7, study of the kinematics of large-scale cosmic structures—galaxies, galaxy clusters, and superclusters—indicates that the bulk of the matter in the universe is invisible (i.e. non-luminous) *dark matter*. The nature of this dark matter is presently unknown. There is so far no direct experimental evidence for detection of individual dark matter particles. It is possible that the proposed massive supersymmetric particles mentioned in Section 1.3 could be candidates.

Present experiments in astrophysics, to be described in the following chapters, also indicate that, although in the past the universal expansion following the Big Bang was indeed slowing down (on account of the restraining pull of gravity), the expansion is now accelerating, this being ascribed to *dark* *energy* which actually exerts a gravitational repulsion. This dark energy exceeds the energy density in all other matter and radiation. Again, the actual nature and origin of this dark energy is presently unknown.

The possibilities of new particles and new interactions also follow from the very successes of the Standard Model, in being able to unify electromagnetic with weak interactions in the electroweak theory. This suggests that it might also be possible to unify strong with electroweak interactions in a so-called grand unified theory (GUT). At present, however, there is no direct experimental support for such higher levels of unification. The reasons for these schemes, and their consequences for the experimental situation are discussed in Chapters 3 and 4. The evidence for higher mass scales, well above anything attainable in the laboratory, may also be suggested by the success of inflationary models of the early universe, described in Chapter 8, and by the evidence for finite neutrino masses, described in Chapters 4 and 9.

Even more ambitious than the proposed unification of strong, weak, and electromagnetic interactions, is the attempt to unify *all* fundamental interactions, including gravity, in so-called *superstring* theory (see Chapter 3). This can produce a unified, renormalizable quantum theory of the fundamental interactions, but only in 10-dimensional space–time (for fermions). One must suppose that all but the normal four dimensions of space and time are 'curled up' to the tiny dimensions of the Planck length (1.12), according to ideas originally advanced by Kaluza and Klein in the 1920s. Unfortunately, apart from the single prediction of spin 2 for the graviton, superstring theories appear so far to make no testable predictions.

1.13 Summary

- Matter is built from elementary fermion constituents, the quarks and leptons, occurring in three 'families'. Each family consists of a quark with charge $+2|e|/3$ and one of $-|e|/3$, a charged lepton with charge $-|e|$, and a neutral lepton (neutrino). The antiparticles of these states have electric charges of opposite sign but are otherwise identical to the particles.
- The observed strongly interacting particles (hadrons) consist of bound quark combinations. Baryons consist of three quarks and mesons of a quark–antiquark pair.
- Quarks and leptons interact via exchange of fundamental bosons, characteristic of four fundamental interactions: strong, electromagnetic, weak, and gravitational. The exchanged bosons are virtual particles.
- Strong interactions occur between quarks and are mediated by gluon exchange; electromagnetic interactions are between all charged particles and are mediated by photon exchange; weak interactions are mediated by $W^{\pm}$ and Z^0 exchange; gravitational interactions via graviton exchange. Photons, gluons, and gravitons are massless. W and Z bosons have masses of 80 and 91 GeV/c^2 respectively.
- The basic boson exchange process can be visualized by diagrams called Feynman diagrams, depicting the exchange of a virtual boson between two interacting fermions. The amplitude for this process is the product of the couplings g_1 and g_2 of the fermions to the exchanged boson,

multiplied by a propagator term which depends on the (free) boson mass M and the momentum transfer q, of the form $1/\left(M^2 - q^2\right)$, where q^2 is a negative quantity. The cross-section for the interaction is the product of the square of the above amplitude and a phase–space factor. It is numerically equal to the reaction rate per target particle per unit incident flux.

- The strength of an interaction is also measured by the decay rate or width of unstable hadronic or leptonic states. If the width is large enough to be measurable, the state is referred to as a resonance.
- Under normal conditions, quarks are confined as combinations in baryons (QQQ) or mesons ($Q\bar{Q}$). At sufficiently high temperatures, with $kT > 0.3$ GeV, it is expected that quarks would no longer be confined, and that hadrons would undergo a phase transition to a quark–gluon plasma.

Problems

A table of physical constants can be found in Appendix A.
More challenging problems are marked with an asterisk.

(1.1) Find the fractional change in total energy (including rest energy and gravitational potential energy) when two equal and isolated point masses M are brought from infinity to a separation R. Calculate what this is when (a) M equals one solar mass and R equals 1 parsec and (b) when M equals the Planck mass and R equals 1 fm.

*(1.2) The bombardment of a proton target by a pion beam of energy 1 GeV/c results in the reaction $\pi^- + p \rightarrow \Lambda + K^0$ with a cross-section of about 1 mb. The K^0 particle is a meson of strangeness $S = +1$, while the Λ is a baryon of strangeness $S = -1$. Write down the above reaction in terms of quark constituents.

Both of the product particles are unstable. One undergoes decay in the mode $\Lambda \rightarrow p + \pi^-$ with a mean lifetime of 10^{-10} s, while the other decays in the mode $K^0 \rightarrow \pi^+ + \pi^-$, also with a lifetime of 10^{-10} s. Both decay rates and interaction cross-sections are proportional to the squares of the coupling constants associated with the interactions responsible. Explain qualitatively how the long lifetimes for the above decays can be reconciled with the large production cross-section. (*Note*: this contrast between the strong production cross-section and long decay lifetime was the reason for calling the Λ and K^0 'strange' particles.)

(1.3) Calculate the energy carried off by the neutrino in the decay of a pion at rest, $\pi^+ \rightarrow \mu^+ + \nu_\mu$. The relevant masses are $m_\pi c^2 = 139$ MeV, $m_\mu c^2 = 106$ MeV, $m_\nu \sim 0$. Pions in a beam of energy 10 GeV decay in flight. What is the maximum and minimum energies of the muons from these decays? (See Chapter 2 for formulae on relativistic transformations.)

(1.4) The Δ^{++} resonance shown in Fig. 1.14 has a full width of $\Gamma = 120$ MeV. What is the mean proper lifetime of this state? How far on average would such a particle, of energy 100 GeV, travel before decaying?

(1.5) The Ω^- is a baryon of mass 1672 MeV/c^2 and strangeness $S = -3$. It decays principally to a Λ baryon of mass 1116 MeV/c^2 and $S = -1$ and a K^- meson of mass 450 MeV/c^2 and $S = -1$. Express the decay process in terms of quark constituents in a Feynman diagram.

State which of the following decay modes are possible for the Ω particle:

(a) $\Omega^- \rightarrow \Xi^0 + \pi^-$ ($m_\Xi = 1315\,\text{MeV/c}^2, S = -2$)
(b) $\Omega^- \rightarrow \Sigma^- + \pi^0$ ($m_\Sigma = 1197\,\text{MeV/c}^2, S = -1$)
(c) $\Omega^- \rightarrow \Lambda^0 + \pi^-$ ($m_\Lambda = 1116\,\text{MeV/c}^2, S = -1$)
(d) $\Omega^- \rightarrow \Sigma^+ + K^- + K^-$ ($m_K = 494\,\text{MeV/c}^2$)

*(1.6) The following decays are all ascribed to the weak interaction, resulting in three final state particles. For each process, the available energy Q in the decay is given together with the decay rate W:

(a) $\tau^+ \rightarrow e^+ + \nu_e + \bar{\nu}_\tau$
$Q = 1775$ MeV
$W = 6.1 \times 10^{11}\text{s}^{-1}$

(b) $\mu^+ \to e^+ + \nu_e + \bar{\nu}_\mu$
$Q = 105$ MeV
$W = 4.6 \times 10^5\ s^{-1}$

(c) $\pi^+ \to \pi^0 + e^+ + \nu_e$
$Q = 4.1$ MeV
$W = 0.39\ s^{-1}$

(d) $^{14}O \to {}^{14}N^* + e^+ + \nu_e$
$Q = 1.8$ MeV
$W = 5.1 \times 10^{-3}\ s^{-1}$

(e) $n \to p + e^- + \nu_e$
$Q = 0.78$ MeV
$W = 1.13 \times 10^{-3}\ s^{-1}$

Using dimensional analysis show that within one or two orders of magnitude, the Q values and decay rates are compatible with the same weak coupling. Comment on any trends with the values of Q.

*(1.7) The cross-section for neutrino–electron scattering via $W-$exchange is given in (1.27b) as $\sigma\,(\nu_e + e \to \nu_e + e) = G_F^2 s/\pi$, where s is the square of the CMS energy. At high energies, deep inelastic neutrino–nucleon scattering can be treated as elastic scattering of a neutrino by a quasi-free quark constituent of the nucleon, the scattered quark then 'fragmenting' into secondary hadrons. Assuming that the struck quark carries on average 25% of the mass of the nucleon, calculate the neutrino–nucleon cross-section in cm^2 as a function of the laboratory neutrino energy in GeV.

*(1.8) In the previous question, the cross-section formula assumes a pointlike interaction specified by the Fermi constant G_F. However at very high energies, the effect of the finite W boson mass in the propagator term (1.9) must be taken into account. Write down an expression for the differential cross-section $d\sigma/dq^2$ for neutrino–electron scattering in this case, based on Equation (1.27b) and the fact that the maximum value of momentum transfer squared is $q^2(\max) = s$. Show that, as $q^2(\max) \to \infty$, the neutrino–electron cross-section tends to a constant, and find its value. At what neutrino energy does the cross-section reach half of its asymptotic value?

(1.9) A resonance of significance for experiments on ultra-high-energy neutrinos in astrophysics is the Glashow resonance:

$$\bar{\nu}_e + e^- \to W^-$$

where $M_W c^2 = 81$ GeV. Assuming that the target electrons in the cosmos are at rest, show that this resonance would be excited for antineutrino energies of around 6400 TeV, and that the peak cross-section would be about 5μb.

(1.10) The charmed meson D^+ (mass 1.87 GeV/c^2) undergoes weak ΔC = 1 decay in the mode $D^+ \to K^0 + l^+ + \nu_l$ where $l = e$ or μ, with a 15% branching ratio. The quark constitution of the charmed meson is $D^+ = c\bar{d}$ and of the kaon is $K^0 = s\bar{d}$, so the decay can be written as the transformation of a charmed to a strange quark (with a $\bar{d}$-quark as 'spectator'):

$$c \to s + l^+ + \nu_l$$

in close analogy with muon decay. Draw the Feynman diagram for c quark decay, and assuming a mass of the c quark of 1.6 GeV/c^2, and neglecting the mass of the decay products, estimate the lifetime of the D meson from that of the muon as given in Problem 1.6.

(1.11) The Σ-baryons are combinations of s, u, and d quarks, with strangeness $S = -1$. The first two entries in the Table on the next page are for the ground-state combination of spin $J = 1/2$ while the third is an excited state of $J = 3/2$. The rest-masses in MeV/c^2 are given in brackets, followed by the quark constitution, the $Q-$value, principal decay mode and lifetime or width. All the decay products have $S = 0$ except for the Λ-baryon, which has $S = -1$.

State which of the fundamental interactions are responsible for the above decays, and from the decay rates estimate the relative values of their coupling strengths.

(1.12) Draw the Feynman diagram representing electron–electron scattering to first order in the coupling constant. If you carefully label the incoming and outgoing electron states, in fact you will find that *two* diagrams are possible. Draw some second-order diagrams involving exchanges of photons and/or electron–positron pairs. Compare the interaction rates with those for the first-order process.

(1.13) The following transitions have Q values and mean lifetimes as indicated:

Transition	Q-value (MeV)	Lifetime (s)
(a) $\mu^+ \to e^+ + \nu_e + \nu_\mu$	105	2.2×10^{-6}
(b) $\mu^- + {}^{12}C \to {}^{12}B + \nu_\mu$	93	2×10^{-6}
(c) $\pi^0 \to 2\gamma$	135	10^{-17}
(d) $\Delta^{++} \to p + \pi^+$	120	10^{-23}

State which interactions are responsible in each case, and estimate the relative coupling strengths from the given quantities.

*(1.14) Calculate the ratio R of the cross-section for $e^+e^- \to Q\bar{Q} \to$ hadrons to that for the reaction $e^+e^- \to \mu^+\mu^-$ in (1.25), as a function of increasing CMS energy up to 20 GeV. Assume the quark masses given in Table 1.4.

At a certain energy, the process $e^+e^- \to \pi^+ + \pi^- + \pi^0$ is observed. Draw a Feynman diagram to illustrate such an event.

Baryon	Quark structure	Q (MeV)	Decay mode	Lifetime or width
Σ^0 (1192)	*uds*	74	$\Lambda\gamma$	$\tau = 7.4 \times 10^{-20}$ s
Σ^+ (1189)	*uus*	187	$p\pi^0, n\pi^+$	$\tau = 8 \times 10^{-11}$ s
Σ^0 (1385)	*uds*	208	$\Lambda\pi^0$	$\Gamma = 36$ MeV

*(1.15) The Δ pion-nucleon resonance (see Fig. 1.14) has a central mass of 1232 MeV/c^2 and spin $J = 3/2$. It decays predominantly into a pion of $J = 0$ plus a nucleon of $J = 1/2$, but also decays in the mode $\Delta \rightarrow n + \gamma$ with a branching ratio of 0.55%. Using Equation (1.36), calculate the peak cross-section for the process $\gamma + p \rightarrow \Delta^+$. The cosmic microwave background consists of photons with a temperature of $T = 2.73$ K and density of 400 cm^{-3}. Estimate the energy that primary cosmic ray protons would require in order to excite the peak of the Δ resonance, in collisions with the microwave background. Assume a photon energy of $2.7kT$ and head-on collisions. What is the mean free path for collision of such protons?

Relativistic transformations and the equivalence principle

2

As a precursor of a discussion of invariance principles and symmetries in Chapter 3, we summarize in this chapter relativistic transformations and Lorentz invariance, the Equivalence Principle, and important solutions of the Einstein field equations of general relativity. These are central to our discussions of cosmology in later chapters. Readers familiar with these topics can skip to Chapter 3.

2.1 Coordinate transformations in special relativity

The *special theory of relativity,* proposed by Einstein in 1905, involves transformations between *inertial frames* (IFs) of reference. An IF is one in which Newton's law of inertia holds: a body in such a frame not acted on by any external force continues in its state of rest or of uniform motion in a straight line. Although an IF is, strictly speaking, an idealized concept, a reference frame far removed from any fields or gravitating masses approximates to such a frame, as does a lift in free fall on Earth. On the scale of experiments in high-energy physics at accelerators, gravitational effects are negligibly small and to all intents and purposes the laboratory can be treated as an IF. However, on the scale of the cosmos, gravity is the most important of the fundamental interactions.

We list here the coordinate transformations, called Lorentz transformations, among IFs in special relativity. These are obtained from two assumptions: that the coordinate transformations should be linear (to agree with the Galilean transformations in the non-relativistic limit); and that the velocity of light c in vacuum should be the same in all IFs (as observed in numerous experiments). The relation between the coordinates x', y', z', t' of an event in an IF Σ' moving with velocity v along the x-axis with respect to an IF Σ, where the coordinates of the event are x, y, z, and t, is then as follows:

$$x' = \gamma (x - vt) \quad y' = y \quad z' = z \quad t' = \gamma \left(t - \frac{vx}{c^2} \right) \tag{2.1}$$

where $\gamma = \left(1 - v^2/c^2\right)^{-1/2}$ is the so-called Lorentz factor. When $v \to 0, \gamma \to 1$, $x' \to (x - vt)$, and $t' \to t$ as in the Galilean transformation. The above transformation also makes the velocity of light invariant: $x'^2+y'^2+z'^2-c^2t'^2 = x^2+y^2+z^2-c^2t^2 = 0$.

According to these transformations, distances in the x-direction as measured in the frame Σ', appear *contracted* when measured in the frame Σ, while time intervals appear *dilated*. First, suppose that a rod aligned with the x-axis is stationary in the frame Σ', where it has a measured length $l' = x'_2 - x'_1$, taken at any time t' in that system. In the frame Σ, the coordinates of the ends of the rod have to be taken *simultaneously at a fixed time t*, by an observer O_1 stationed half way between, and receiving simultaneous signals from, observers O_2 and O_3 at each end of the rod, as it moves past (see Fig. 2.1). Then according to (2.1) these observers will record

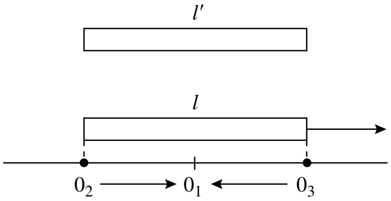


Fig. 2.1

$$x_1 = \frac{x'_1}{\gamma} + vt$$

$$x_2 = \frac{x'_2}{\gamma} + vt$$

and hence the observers in Σ measure the 'contracted' length to be

$$l = x_2 - x_1 = \frac{(x'_2 - x'_1)}{\gamma} = \frac{l'}{\gamma} \tag{2.2}$$

Equally, observers in frame Σ' would measure the length of a rod, of length x at rest in the Σ frame, to be $l' = l/\gamma$. This does *not* conflict with (2.2). In either case, it is the moving rod which appears to be contracted in length, and the two situations are asymmetric, since three observers are required to measure the moving rod, but only one is needed in the frame where the rod is at rest. Of course, the rod does not actually contract. It is simply that the measurement of length differs for observers in relative motion.

Second, suppose there is a clock at rest in the moving frame Σ'. It will be located at some fixed coordinate x'. Then let t'_1 and t'_2 be two times recorded by this clock in the frame Σ'. The times recorded by the observer in the frame Σ will therefore be, from the inverse transformation to (2.1)

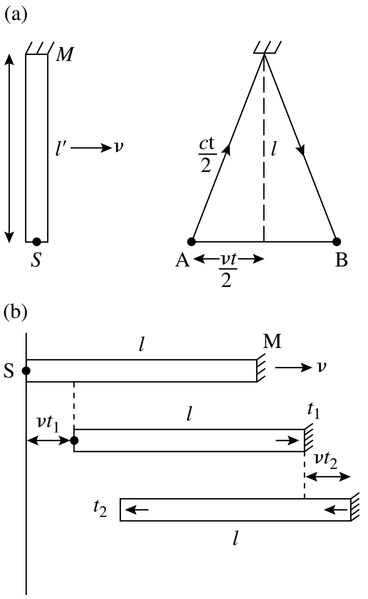


Fig. 2.2

$$t_1 = \gamma\left(t'_1 + \frac{vx'}{c^2}\right)$$

$$t_2 = \gamma\left(t'_2 + \frac{vx'}{c^2}\right)$$

so that in this case

$$t = t_2 - t_1 = \gamma\left(t'_2 - t'_1\right) = \gamma t' \tag{2.3}$$

and the timescale of the moving clock appears dilated when compared with an identical clock in the rest-frame of the stationary observer.

The above relations can be understood from simple geometrical constructions, using a rod, a light source, and a mirror, as shown in Fig. 2.2(a) and (b) and Example 2.1.

Example 2.1 *Verify the above transformations, by taking the case of a moving rod carrying a light source at one end and a mirror at the other, the rod being aligned (a) normal and (b) parallel to the direction of motion.*

Refer first to Fig. 2.2(a). The rod with a pulsed light source S at one end and a mirror M at the other, is at rest and of length l' in the Σ' frame, which is moving at velocity v along the x-axis of the frame Σ, the rod being aligned with the y-axis. The time for the light pulse to travel to the mirror and back is clearly $t' = 2l'/c$, as measured in the frame Σ'. However, in the frame Σ, the return journey time is t, during which time the rod has moved a distance vt along the x-axis. Since transverse distances are the same in the two frames ($l = l'$), the right-angled triangle gives

$$\left(\frac{ct}{2}\right)^2 = \left(\frac{ct'}{2}\right)^2 + \left(\frac{vt}{2}\right)^2$$

$$\text{or} \qquad t = \frac{t'}{\sqrt{\left(1 - v^2/c^2\right)}} = \gamma t' \tag{2.4}$$

as before. In the second arrangement, in Fig. 2.2(b), the rod is aligned in the x-direction. The time for the light pulse to travel to the mirror and back is again $t' = 2l'/c$ in the Σ' frame. However, in the Σ frame the time t_1 for the pulse to reach the mirror is longer because the mirror has moved, and is given by $ct_1 = l + vt_1$, or $t_1 = l/(c - v)$. The return signal from the mirror reaches the source after time t_2, where $ct_2 = l - vt_2$, or $t_2 = l/(c + v)$. Hence the total time in the Σ system is

$$t = t_1 + t_2 = \frac{2l}{c\left(1 - v^2/c^2\right)} \tag{2.5}$$

However, from (2.4) we know that

$$t = \frac{t'}{\sqrt{\left(1 - v^2/c^2\right)}} = \frac{2l'}{c\sqrt{\left(1 - v^2/c^2\right)}} \tag{2.6}$$

Comparing the last two expressions we obtain the result (2.2) for the apparent contraction

$$l = \frac{l'}{\gamma} \tag{2.7}$$

2.2 Invariant intervals and four-vectors

We have seen that the measured values of time and space intervals for the same event differ in different reference frames. However, the goal of relativity theory is to formulate the equations of physics in a way that is invariant in all reference frames. The *invariant interval* or *line element* is made up of the squares of the coordinate differences for two events and is defined by

$$\begin{aligned} \mathrm{d}s^2 &= c^2\mathrm{d}t'^2 - \mathrm{d}x'^2 - \mathrm{d}y'^2 - \mathrm{d}z'^2 \\ &= c^2\mathrm{d}t^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2 \\ &= c^2\mathrm{d}\tau^2 \end{aligned} \tag{2.8}$$

It has the same value in all IFs, as is easily demonstrated by substitution in (2.1). The interval is therefore invariant under Lorentz transformations between IFs. Since it involves as components three space and one time coordinate, ds is referred to as a *four-vector*. Obviously, it is desirable to describe physical equations in terms of four-vector quantities, so that they hold in all IFs. The invariance of four-vectors under Lorentz transformations between IFs is analogous to the invariance of the lengths of three-vectors under rotations or translations in three-dimensional space. In the third line in (2.8) above, τ refers to the *proper time,* that is, the time on a clock fixed in the IF (for which $\mathrm{d}x = \mathrm{d}y = \mathrm{d}z = 0$). The interval d$s$ is referred to as *timelike, null*, or *spacelike* according to $\mathrm{d}s^2$ being positive, zero, or negative.

We note that if the coordinate increments refer to the passage of a light-ray, $\mathrm{d}s^2 = 0$, using Pythagoras's theorem and the fact that in IFs, light travels in straight lines. In the case of non-IFs, that is, reference frames accelerating with respect to IFs, light does not travel in straight lines, and space/time is non-Euclidean or 'curved', as described below. However, $\mathrm{d}s^2 = 0$ again for a light-ray in such a non-IF, since it is always possible to define IFs (with $\mathrm{d}s^2 = 0$) which are instantaneously co-moving with—and thus have the same transformation properties as—the accelerated frame (AF). Hence the interval in the AF can be subdivided into a succession of tiny intervals in co-moving IFs, for each of which $\mathrm{d}s^2 = 0$. So the zero value of the interval is a general property of photons (or any other massless particles) in any reference frame.

In the description given before we used rectangular coordinates, but we can also set the interval in terms of spherical coordinates r, θ, φ, where r is the radial coordinate, θ is the polar angle, and φ is the azimuthal angle about the z-axis. We can write

$$\mathrm{d}s^2 = c^2\mathrm{d}t^2 - \mathrm{d}r^2 - r^2\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\varphi^2\right) \tag{2.9}$$

that is, dx is replaced by dr, dy by $r\mathrm{d}\theta$, and dz by $r\sin\theta\,\mathrm{d}\varphi$.

The *general theory of relativity* proposed by Einstein in 1915 is concerned with providing an invariant description of physical phenomena in all conceivable reference frames, including those in accelerated motion with respect to IFs, the acceleration being provided by gravitational fields. Einstein's field equations of general relativity were based on the very important *equivalence principle*, which we now discuss.

2.3 The equivalence principle: clocks in gravitational fields

Suppose that an observer, initially at rest in an IF Σ, is given a *small* acceleration a in the x-direction. The space coordinates he or she records, for an event with coordinates x, y, z, and t in Σ will be, according to Newtonian mechanics

$$x' = x - \frac{1}{2}at^2, \quad y' = y, \quad z' = z$$

With $x = x' + 1/2at^2$ and $\mathrm{d}x = \left(\partial x/\partial x'\right) \mathrm{d}x' + (\partial x/\partial t)\ \mathrm{d}t = \mathrm{d}x' + at\,\mathrm{d}t$, the invariant interval (2.8) is

$$\begin{aligned} \mathrm{d}s^2 &= c^2\mathrm{d}t^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2 \\ &= \left(c^2 - a^2t^2\right) \mathrm{d}t^2 - 2at\,\mathrm{d}x'\mathrm{d}t - \mathrm{d}x'^2 - \mathrm{d}y'^2 - \mathrm{d}z'^2 \end{aligned} \tag{2.10}$$

The second line refers to spatial coordinates measured in the AF Σ'. The time $\mathrm{d}t'$ elapsed on a clock fixed in this accelerating frame of reference, that is, for which $\mathrm{d}x' = \mathrm{d}y' = \mathrm{d}z' = 0$, will be given by $\mathrm{d}s^2 = c^2\mathrm{d}t'^2$ and hence

$$\mathrm{d}t'^2 = \left(1 - a^2t^2/c^2\right) \mathrm{d}t^2 \tag{2.11}$$

The instantaneous velocity of the accelerating clock measured in Σ is $v = at$ so that the interval $\mathrm{d}t'$ of proper time measured on this clock, as compared with the value $\mathrm{d}t$ measured on an identical clock at rest in Σ is also given by

$$\mathrm{d}t'^2 = \left(1 - v^2/c^2\right) \mathrm{d}t^2 \tag{2.12}$$

which is the usual formula for time dilation in (2.3). According to the accelerated observer, the time intervals $\mathrm{d}t$ on the clock in the frame Σ are dilated in comparison with intervals $\mathrm{d}t'$ on his own clock. The interval $\mathrm{d}t'$ here is the same as would be measured on an identical clock at rest in an *IF* Σ'' which is instantaneously co-moving with the accelerated clock and has velocity v with respect to the frame Σ. The distance which the accelerated clock has moved after time t is $H = (1/2)at^2$, so that $a^2t^2 = 2aH$ and (2.11) can be written as

$$\mathrm{d}t'^2 = \left(1 - \frac{2aH}{c^2}\right) \mathrm{d}t^2 \tag{2.13}$$

Einstein's *Principle of Equivalence* states that a frame Σ' accelerating with respect to an IF Σ is exactly equivalent to a system at rest in Σ but subject to a *homogeneous* gravitational field. Note the word 'homogeneous'. The force of gravity on the Earth is not homogeneous, as it points towards the Earth's centre, and is therefore in different directions in different places. A free-falling lift in Sydney is accelerated in a different direction from one in London. But in one *localized* position, of extent very small compared with the Earth's radius, the gravitational field is almost homogeneous. Indeed, *tidal forces* arise because of slight differences in direction or magnitude of the gravitational force over finite distances (see Fig. 2.3 and Example 2.2).

Example 2.2 *Calculate the height of the diurnal mid-ocean tide, due to the Moon's gravitational pull. Assume that the solid Earth itself is completely rigid. Compare the tidal effects of the Sun and the Moon.*

Figure 2.3 shows (not to scale) the Earth, centre O, mass M_e, and radius R, and the moon, mass M_m at distance D from O. The potential at point P

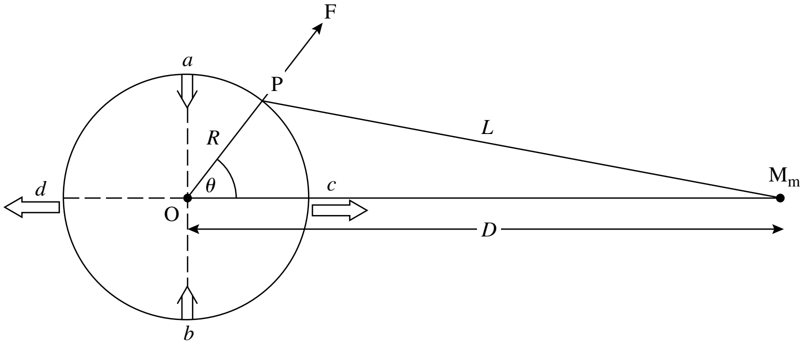


Fig. 2.3 Tidal forces, shown here by the broad arrows, arise because the gravitational attraction of a distant body—the Moon in this example—varies over a finite distance on the Earth's surface. See Example 2.2.

due to lunar gravity is (with G as the gravitational constant):

$$V = \frac{GM_m}{L} = GM_m \left(D^2 + R^2 - 2RD\,\cos\theta\right)^{-1/2}$$
$$= \left(\frac{GM_m}{D}\right)\left[1 - \left(\frac{R^2}{2D^2}\right) + \left(\frac{R}{D}\right)\cos\theta + \left(\frac{3R^2}{2D^2}\right)\cos^2\theta + \cdots\right]$$

where the second line is the result of a binomial expansion, and we neglect terms in $(R/D)^3$ or higher powers. The radial force at P per unit mass is then

$$F = -\frac{\partial V}{\partial R} = -\left(\frac{GM_m}{D^2}\right)\cos\theta + \left(\frac{GM_mR}{D^3}\right)\left(3\,\cos^2\theta - 1\right)$$

The first term on the right is just the radial component of the force GM_m/D^2 per unit mass experienced by the entire Earth due to lunar gravity. The second term is the tidal force. For $\theta = 0$ or π, it is radially outwards (i.e. a rising tide) shown at points c and d in Fig. 2.3, while for $\theta = \pi/2$ or $3\pi/2$, it is radially inwards (i.e. a falling tide) at points a and b. The magnitude of the tidal force is of order GM_mR/D^3, to be compared with the Earth's gravitational force on unit mass at the Earth's surface of GM_e/R^2. If the tidal height is h, then the reduction in the Earth's gravitational force over this distance must be just equal to the lunar tidal force:

$$h\,\frac{\partial\left(GM_e/R^2\right)}{\partial R} = \frac{2GM_mR}{D^3}$$

or

$$\frac{h}{R} = \left(\frac{M_m}{M_e}\right)\left(\frac{R}{D}\right)^3$$

Inserting the constants $M_e = 5.98 \times 10^{24}$ kg, $M_m = 7.34 \times 10^{22}$ kg, $D = 3.84 \times 10^8$ m, $R = 6.37 \times 10^6$ m, one obtains $h = 0.4$m. This is a slight underestimate, since the Earth is not completely rigid and bulges a little under the lunar force. (The tides observed around shallow seas in coastal areas or estuaries are of course very much higher, typically 10 or 20 times the height in the deep ocean.)

The Sun's gravitational force on the Earth is enormous compared with that of the Moon, but the tides depend on the derivative of the force, and using the constants given in Appendix A it is easy to show that the solar tide is less than half that due to the Moon.

That a homogeneous gravitational field is equivalent to a uniform acceleration with respect to an IF comes about because of the equivalence of gravitational and inertial mass. The inertial mass of a body is defined as the ratio of the force F_I applied to the acceleration g produced, that is,

$$F_{\mathrm{I}} = M_{\mathrm{I}} g$$

The gravitational mass is defined by the force on the body in a gravitational field, due, for example, to a point mass M at distance r:

$$F_{\mathrm{G}} = M_{\mathrm{G}} \left(\frac{GM}{r^2} \right)$$

If F_I is the gravitational force F_{G} it follows that the 'gravitational acceleration' is

$$g = \left(\frac{M_{\mathrm{G}}}{M_{\mathrm{I}}} \right) \frac{GM}{r^2} \tag{2.14}$$

so that g will be the same for all bodies provided they have the same ratio of gravitational to inertial mass. The principle of equivalence has been checked experimentally using very precise torsion balance experiments. These have a long history, starting from the pioneer experiments by Baron Eötvos in Budapest in the 1920s. Figure 2.4 shows the principle of such a torsion balance experiment. A body A at the Earth's surface at latitude λ is subject to two forces; the gravitational force F_G proportional to the gravitational mass M_G along the line AB towards the Earth's centre, and a centripetal force F_I proportional to the inertial mass M_I along AC arising from the Earth's rotation. If a body is suspended by a string, the string will lie along the resultant AD of these forces, the angle θ to the local vertical depending on the ratio $R = M_{\mathrm{I}}/M_{\mathrm{G}}$.

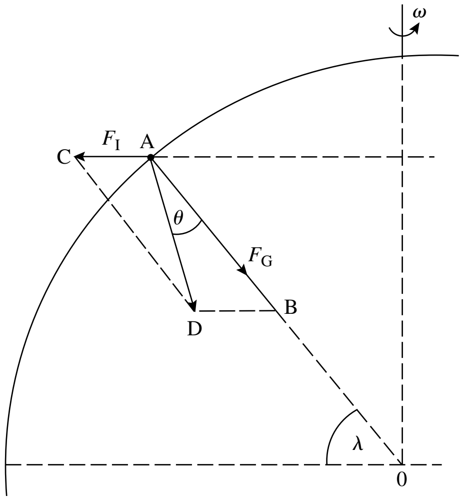


Fig. 2.4 The force on a body at latitude λ at sea-level is the resultant AD of the gravitational force F_{G}, proportional to the gravitational mass M_{G}, and the centripetal force F_{I}, proportional to the inertial mass M_{I}.

According to the Equivalence Principle, R should be the same for all bodies of whatever material (and if so we define units so that $R = 1$, and simply refer to 'the mass' of the body). To test this, two bodies of equal masses but different materials are suspended from either end of a horizontal beam, itself hanging from a torsion fibre. If R is different for different materials, so also will be the value of θ, and the result is a net couple on the suspension, which will change sign on rotating the entire apparatus through 180°. For a suspension with a suitably long natural period of oscillation, a higher sensitivity can be obtained by utilizing the gravitational field of the Sun, rather than the Earth. Then one searches for a 12-h oscillation period of the torsion system, as the magnitude of the centripetal force due to the Earth's rotation changes sign relative to the Sun's gravitational field. By applying this method, Braginsky and Panov (1972), using masses of platinum and aluminium, set a limit on the possible difference in R for these two substances of $\Delta R/R < 10^{-12}$.

The above experiment is sensitive to the gravitational field at large distances, of the order of an astronomical unit. A different type of torsion arrangement, what one might call a 'table top' experiment, was performed by Gundlach *et al.* (1997), and was sensitive to any possible deviations from the Equivalence Principle at distances as small as a centimetre. It compared the accelerations of copper and lead masses towards a massive attractor consisting of 3 tons of uranium. The experiment found a fractional difference in this case of $\Delta R/R < 10^{-8}$. To summarize therefore, the Equivalence Principle seems to be well satisfied.

According to (2.14), all bodies, irrespective of mass, will therefore have the same gravitational acceleration—first demonstrated by the (probably apocryphal) story of Galileo at the Leaning Tower of Pisa. This means that, if a person in a freely falling lift holds Newton's proverbial apple at arm's length and then releases it, it will remain exactly where it is, relative to the frame of the lift. So, according to Newton's law of inertia, a local region in a freely falling lift is indeed an IF, since the apple remains in its state of rest in that system. As stated before, this is strictly true only in a small region of space where the gravitational field is essentially homogeneous. In an actual lift of finite extent on Earth, apples placed apart, on either side of the lift, will of course experience a tidal force and gradually move towards each other (see Fig. 2.3). The contrasting situations for a weight suspended by a spring in a box, which is placed first in an IF, then in a gravitational field, then in an AF, and finally in a free-fall frame are depicted in Fig. 2.5. Although in the above, we have considered only mechanics, the (so-called) *strong equivalence principle* asserts that for *all* the fundamental interactions, a freely falling frame in a (homogeneous) gravitational field is an IF.

From the above discussion and equations (2.12) and (2.13), we see that we can replace the *accelerated, moving* clock by an identical, *stationary* clock in a gravitational field providing a gravitational acceleration a, so that

$$\mathrm{d}t'^2 = \left(1 + \frac{2\Delta\Phi}{c^2}\right)\mathrm{d}t^2 \tag{2.15}$$

where $\mathrm{d}t'$ is the time interval on the clock in the field, $\mathrm{d}t$ is that on an identical clock in an IF remote from any gravitational field, and $\Delta\Phi = -aH$ is the difference in gravitational potential. For the remote clock, $\Phi = 0$, while for that

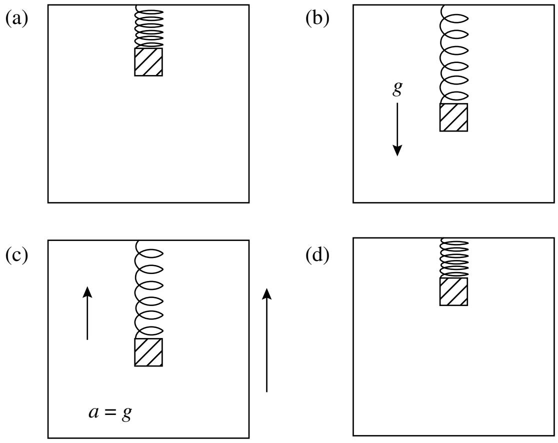


Fig. 2.5 A mass suspended from a spring attached to the roof of a closed box, in four contrasting cases. In (a) the box is in an inertial frame, and the spring is not extended. In (b) the box is in a gravitational field, equivalent to an acceleration g, with the spring extended. In (c) the box is accelerated with an acceleration $a = g$ with respect to an inertial frame, again with the spring extended. Finally in (d) the box is in free fall in a homogeneous gravitational field, with no extension of the spring. Because of the equivalence principle, an observer inside the box could not distinguish (b) from (c) or (d) from (a).

in the field, $\Phi < 0$. Thus a clock at low (negative) gravitational potential, such as one at sea-level, should run slower than an identical clock at a higher (less negative) potential on a mountain top. This predicted gravitational shift was verified experimentally by Pound and Snider (1964). In their experiment, the very small (10^{-15}) increase in frequency f of ^{57}Fe γ-rays falling down a vertical 22-m tube was measured by means of an ^{57}Fe absorber at the bottom, utilizing the Mossbauer effect. The photons from the emitter at the higher potential are 'blue-shifted' compared with the absorption frequency at the lower potential, and this was compensated using the Doppler effect, by slowly moving the absorber downwards at the appropriate velocity $v/c = \Delta f/f \sim 10^{-15}$. Since that time, atomic clocks have been carried on aircraft to directly verify the above formula by comparing with similar clocks at ground level.

Relative to our remote clock at $\Phi = 0$, a clock in the field of a point mass M at distance r is at a potential $\Phi = -GM/r$ and

$$\mathrm{d}t'^2 = \left[1 - \frac{2GM}{(rc^2)}\right] \mathrm{d}t^2 \tag{2.16}$$

The analysis here has assumed small values of acceleration, that is, $\Delta\Phi \ll c^2$. It happens, however, that (2.15) and (2.16) are correct even for strong fields, and give the same results as the full analysis using the general theory of relativity.

2.4 General relativity

A full treatment of general relativity is lengthy and outside the scope of this text, and for such a treatment the reader is referred, for example, to the accompanying book in the Oxford Master Series *Relativity, Gravitation and Cosmology* by Ta-Pei Cheng. Here, we just show the form of the Einstein field equations and discuss two important solutions, which we shall be using in later chapters.

In the special theory, equations of physics valid in all IFs could be expressed in terms of scalar and vector quantities (i.e. tensors of zero rank and first rank, respectively). In the general theory, however, the quantities occurring in the (so-called) covariant physical equations, valid in all reference frames, must be expressed as second-rank tensors. If we write the coordinates in (2.1) in a different notation, as $ct = x^0, x = x^1, y = x^2, z = x^3$ then (2.8) can be set in the form of the space–time *metric*

$$\mathrm{d}s^2 = \sum g_{\mu\nu}\, \mathrm{d}x^\mu\, \mathrm{d}x^\nu = \sum \mathrm{d}x^\mu \mathrm{d}x_\mu \tag{2.17}$$

where the summation is over $\mu, \nu = 0, 1, 2, 3$, and $g_{\mu\nu}$ is a 4×4 matrix called the *metric tensor*.

Here, the coordinates have been labelled with upper (contravariant) indices, and the metric tensor with lower (covariant) indices, according to how they transform under a change of coordinate system. Invariant scalars are always a product of covariant and contravariant quantities. For coordinate frames in general, including those accelerating with respect to IFs, the elements of $g_{\mu\nu}$ will be a function of the space–time coordinates x^μ. However, for IFs only, it has a simple form with constant diagonal elements and all off-diagonal elements equal to zero:

$$g_{00} = +1, \quad g_{11} = g_{22} = g_{33} = -1; \quad g_{\mu\nu} = 0 \quad \text{for } \mu \neq \nu. \tag{2.18}$$

or set in matrix form

$$g_{\mu\nu} = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{vmatrix}$$

in conformity with (2.8).

In general relativity, the metric tensor, as the name implies, describes the geometrical properties of space/time, and in particular how it differs from the 'flat' so-called Minkowski metric of special relativity. Einstein had indeed interpreted gravitational effects as due to the geometry of space, which changes in the presence of masses that introduce curvature or 'warping'. The most important quantity in describing deviations from flat space is the Riemann curvature tensor, which is a function of the derivatives of $g_{\mu\nu}$. From this one can derive quantities called the Ricci tensor $\mathfrak{R}_{\mu\nu}$ and the Ricci scalar $\mathfrak{R} = g^{\mu\nu}\mathfrak{R}_{\mu\nu}$. These enter into the expression for the Einstein tensor, $G_{\mu\nu} = \mathfrak{R}_{\mu\nu} - (\mathfrak{R}/2)\; g_{\mu\nu}$. This tensor is symmetric ($G_{\mu\nu} = G_{\nu\mu}$) and has zero divergence, and Einstein's leap of genius was to propose that it should be proportional to the energy momentum tensor $T_{\mu\nu}$, which, due to conservation of energy and momentum, is also symmetric and divergenceless. He deduced the constant of proportionality from Newton's law of gravitation in the non-relativistic, weak field limit. Then the Einstein field equations become, with G as Newton's gravitational constant:

$$G_{\mu\nu} = -8\pi G\, \frac{T_{\mu\nu}}{c^4} \tag{2.19}$$

There are in all $4 \times 4 = 16$ equations, but because of the symmetry of $G_{\mu\nu}$, this reduces to 10, of which only 6 are independent. In the static limit, the relevant $\mu = \nu = 0$ components are $G_{00} = -2\nabla^2\Phi/c^2$ and $T_{00} = \rho c^2$, where Φ is the gravitational potential and ρ is the matter density. In this limit the above set of equations then become Poisson's equation of Newtonian gravity

$$\nabla^2\Phi = 4\pi G\rho \tag{2.20}$$

For a spherically symmetric potential, $\nabla^2\Phi = \left(1/r^2\right)\left[\partial\left(r^2\partial\Phi/\partial r\right)/\partial r\right]$, which upon integration gives

$$\Phi\left(r\right) = \frac{2\pi G\rho r^2}{3}$$

and for the field due to a point mass M at the origin $r = 0$, Newton's inverse square law becomes

$$F\left(r\right) = \frac{\partial\Phi}{\partial r} = \frac{GM}{r^2}$$

The foregoing paragraphs have been inserted simply to show the form of the Einstein field equations. Two important solutions are quoted in the following paragraphs.

2.5 The Schwarzschild line element, Schwarzschild radius, and black holes

A very important solution of the Einstein field equations of general relativity is that obtained by Schwarzschild in 1916, for the metric in the neighbourhood of a spherically symmetric distribution of total mass M, far removed from other gravitating masses. In this case the elements of the metric tensor are no longer constant as in (2.18), but functions of the coordinates. The Schwarzschild line element has the form

$$\mathrm{d}s^2 = \left[1 - \frac{2GM}{\left(rc^2\right)}\right]c^2\mathrm{d}t^2 - \left[1 - \frac{2GM}{\left(rc^2\right)}\right]^{-1}\mathrm{d}r^2 - r^2\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\varphi^2\right) \tag{2.21}$$

where the spherical polar coordinates t, r, θ, and φ are those measured by an observer in an IF, that is, a frame in free fall with respect to the mass M. If we set $\mathrm{d}r = d\theta = \mathrm{d}\varphi = 0$, the proper time interval $\mathrm{d}\tau$ is given by $\mathrm{d}\tau^2 = \mathrm{d}s^2/c^2$, as measured on a clock at rest in the gravitational field of the mass M. From the first term on the right-hand side of the above equation we see that this is reduced in comparison with the value dt on an identical clock in the (free-fall) IF, according to

$$\mathrm{d}\tau = \sqrt{\left[1 - \frac{2GM}{\left(rc^2\right)}\right]}\cdot\mathrm{d}t \tag{2.22}$$

as indicated already in (2.16). A clock in a gravitational field runs slow! When the clock is placed at the radial coordinate $r = 2GM/c^2, \mathrm{d}\tau = 0$, so that the time on the local clock appears to be *frozen*. The quantity

$$r_s = \frac{2GM}{c^2} \tag{2.23}$$

is called the *Schwarzschild radius* of the mass M. What does this mean? Consider a particle with velocity v approaching the mass M in a radial direction.

The inward acceleration it experiences is

$$\frac{\mathrm{d}v}{\mathrm{d}t} = -v\frac{\mathrm{d}v}{\mathrm{d}r} = \frac{GM}{r^2}$$

On integrating, assuming $v = 0$ at $r = \infty$, we obtain for the velocity at radius r

$$v^2 = \frac{2GM}{r}$$

Reversing the path of the particle, we see that v is the escape velocity at r. Therefore the meaning of the Schwarzschild radius in (2.23) is that the escape velocity at this radius is $v = c$, and hence no particle, not even a photon, can escape from inside it. Thus it appears as a *black hole*. We also note that the frequency of light at the Schwarzschild radius goes to zero (i.e. the gravitational redshift is infinite) as demonstrated by the Doppler shift formula (2.36) below. A discussion of the Schwarzschild radius and experimental evidence for black holes is given in Chapters 9 and 10.

The formula (2.21) can be used to calculate the deflection of light-rays by the gravitational field of the mass M, and it provided an early verification of Einstein's general theory of relativity. With the help of special relativity and the equivalence principle, we may note here that the form of the Schwarzschild solution can in fact be understood from purely heuristic arguments. In fact, the equivalence principle was already used by Einstein in arriving at his general theory. The treatment here follows that presented by Adler, Balzin, and Schiffer (1965) by determining the factors which connect the expression for the interval in an IF to that in a frame at rest in the gravitational field of the isolated point mass M (at $r = 0$), which we refer to as the AF. In the IF, the interval in spherical coordinates will be given by the special relativity formula (2.9):

$$\mathrm{d}s^2 = c^2\mathrm{d}t^2 - \mathrm{d}r^2 - r^2\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\varphi^2\right) \tag{2.24}$$

where for a photon, $\mathrm{d}s^2 = 0$. According to the equivalence principle, the length of a standard rod at rest in the AF, as measured by the free-falling IF observer in the (radial) direction of the acceleration, appears *contracted* relative to an identical rod at rest in the IF. The (length)^2 is reduced by a factor $\left(1 - v^2/c^2\right) = \left(1 - 2GM/rc^2\right)$, where v is the instantaneous velocity at r of the free-fall IF with respect to M. On the other hand, the time interval on a clock at rest in the AF appears *dilated* when measured by the IF observer, as compared with an identical clock at rest in the IF, because a clock in a gravitational field runs slow. The (time interval)^2 is increased by a factor $\left(1 - 2GM/rc^2\right)^{-1}$ as in (2.16). The spherical symmetry of the problem means that these factors depend only on the radius vector r. From the values of t, r, θ, and φ and their increments, now measured in the presence of the gravitational field by the IF observer, we can therefore obtain the expression for the interval by dividing by the above factors. So (2.24) becomes

$$\mathrm{d}s^2 = c^2\left(1 - \frac{2GM}{rc^2}\right)\mathrm{d}t^2 - \left(1 - \frac{2GM}{rc^2}\right)^{-1}\mathrm{d}r^2 - r^2\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\varphi^2\right)$$

in conformity with the Schwarzschild line element (2.21). Note that we have *not* derived this transformation from first principles, but merely shown that (2.21) is consistent with what is known from special relativity and the equivalence principle.

2.6 The gravitational deflection of light by a point mass (the Einstein star shift)

Consider the angular deflection of a light beam passing near an isolated point mass M. The *effective* velocity of light, c' is obtained from (2.21) by setting $ds^2 = 0$. If we only want the r dependence, $c'(r)$, then fixing θ and φ we get

$$c'^2 = \frac{dr^2}{dt^2} = c^2\left(1 - \frac{2GM}{rc^2}\right)^2$$

and

$$c'(r) = c\left(1 - \frac{2GM}{rc^2}\right) \tag{2.25}$$

What is meant by 'effective velocity of light' here is that the value of dr/dt differs from that for light in an IF because of the way that distances and times transform in the presence of a gravitational field. The velocity of light passing near M *appears* to be reduced. The effect of the gravitational field is the same as that of introducing an index of refraction, $n = c/c'$, and the deflection can be calculated in the same way.

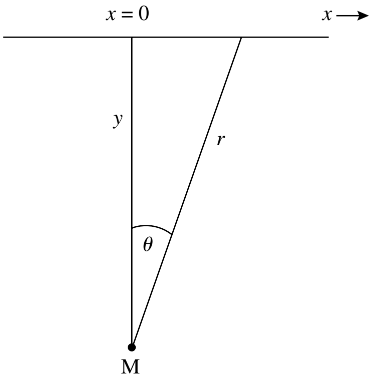


Fig. 2.6

In the scattering of light by the gravitational field of the point mass, assume the beam travels along the x-axis, so we need to find c' as a function of x, for a given impact parameter y along the vertical y-axis—see Fig. 2.6. There is no dependence on azimuth φ, so from (2.21) we get (with $GM \ll rc^2$):

$$\begin{aligned}
ds^2 &= c^2\left(1 - \frac{2GM}{rc^2}\right)dt^2 - \left(1 + \frac{2GM}{rc^2}\right)\left(\frac{dr^2}{dx^2}\right)dx^2 - r^2\left(\frac{d\theta^2}{dx^2}\right)dx^2 \\
&= c^2\left(1 - \frac{2GM}{rc^2}\right)dt^2 - dx^2\left[\left(1 + \frac{2GM}{rc^2}\right)\left(\frac{x^2}{r^2}\right) + \frac{r^2y^2}{r^4}\right] \\
&= c^2\left(1 - \frac{2GM}{rc^2}\right)dt^2 - dx^2\left(1 + \frac{2GMx^2}{c^2r^3}\right)
\end{aligned}$$

In this case, the effective light velocity is given by

$$c'^2 = \left(\frac{dx}{dt}\right)^2 = c^2\left[1 - \frac{2GM\left(1 + x^2/r^2\right)}{rc^2}\right] \tag{2.26}$$

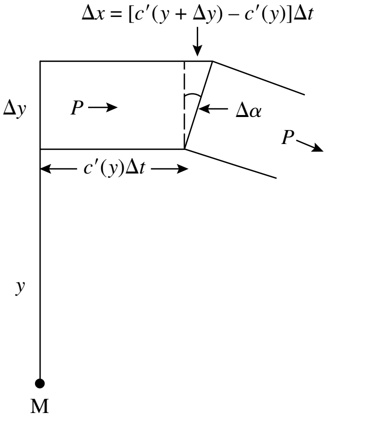


Fig. 2.7

The deflection of the wave vector P over a short time interval Δt is (see Fig. 2.7)

$$\Delta\alpha = \frac{\Delta x}{\Delta y} = \left[c'\left(y + \Delta y\right) - c'\left(y\right)\right]\frac{\Delta t}{\Delta y}$$

$$\text{so}\qquad d\alpha = \left(\frac{dc'}{dy}\right)dt$$

$$\text{and}\qquad \frac{d\alpha}{dx} = \left(\frac{1}{c'}\right)\frac{dc'}{dy}$$

Since the difference between c' and c is very small, we can set $c' = c$ in the denominator. From (2.26) we therefore find

$$\left(\frac{1}{c'}\right)\frac{\mathrm{d}c'}{\mathrm{d}y} = \left(\frac{GM}{r^2c^2}\right)\left[\frac{y}{r} + \frac{3x^2y}{r^3}\right]$$

In the scattering process, the value of y is essentially constant and equal to the impact parameter, $y = b$, say. Then

$$\frac{\mathrm{d}\alpha}{\mathrm{d}x} = \left(\frac{1}{c}\right)\frac{\mathrm{d}c'}{\mathrm{d}y} = \left(\frac{GM}{c^2r^2}\right)\left[\frac{b}{r} + \frac{3x^2b}{r^3}\right] \tag{2.27}$$

With $r = b\sec\theta, x = b\tan\theta, \mathrm{d}x = b\sec^2\theta\mathrm{d}\theta$ one obtains for the deflection of light associated with the passage of the beam through the angular interval $\mathrm{d}\theta$:

$$\mathrm{d}\alpha = \left(\frac{GM}{c^2b^2}\right)\left[\cos^3\theta + 3\sin^2\theta\cos^3\theta\right] b\sec^2\theta\,\mathrm{d}\theta$$

$$\text{and} \qquad \alpha = \left(\frac{GM}{c^2b}\right)\int\left(1 + 3\sin^2\theta\right)d\,(\sin\theta)$$

Inserting the limits of the integral, $\theta = \pi/2$ and $\theta = -\pi/2$ we obtain for the deflection

$$\alpha = \frac{4GM}{c^2b} \tag{2.28}$$

As is well known, the deflection of light from stars close to the Sun was first measured, and the predicted value of 1.75 arcsec (8.48 µrad) from (2.28), was verified by Eddington and others in the 1919 eclipse expedition to Principe and Sobral. It was a first confirmation of Einstein's general theory of relativity.

2.7 Shapiro time delay

The gravitational field due to a point mass M introduces, as well as an angular deflection, a time delay, on light (or any electromagnetic pulse) passing nearby. It is called the Shapiro delay, named after the physicist who was first to realize its existence. It arises because of the 'curvature' or 'warping' of space, which sends photons on a slightly longer path. This time delay was first observed by bouncing radar pulses from Mercury and Venus and observing the delay of the return pulse as the line of sight to these planets moved near to the Sun. To calculate the effect, we neglect the tiny angular deflection in (2.28), so that from (2.26) the increment of transit time over the path element dx becomes (for $GM \ll rc^2$)

$$\mathrm{d}t = \frac{\mathrm{d}x}{c'} = \left(\frac{\mathrm{d}x}{c}\right)\left[1 + \left(\frac{GM}{rc^2}\right)\left(1 + \frac{x^2}{r^2}\right)\right]$$

and the total transit time, from $x = 0$ to $x = X$ is (changing variables to $x = b\tan\theta$)

$$t = \left(\frac{X}{c}\right) + \left(\frac{GM}{c^3}\right)\int \mathrm{d}\theta\frac{\left(1 + \sin^2\theta\right)}{\cos\theta}$$

where θ varies from 0 at $x = 0$ to $\theta_m = \tan^{-1}(X/b)$ for $x = X$. Thus the integrated time delay over the interval $x = 0$ to $x = X$ becomes

$$\Delta t = \left(\frac{GM}{c^3}\right)\left[2\ln\left\{\tan\left(\frac{\theta_m}{2} + \frac{\pi}{4}\right)\right\} - \sin\theta_m\right]$$
$$\approx \left(\frac{GM}{c^3}\right)\left[2\ln\left(\frac{2X}{b}\right) - 1\right] \tag{2.29}$$

where the approximate form, obtained by expanding the circular functions, is good to better than 0.2% for $X > 10b$. In a practical case, we can take for M the solar mass and for the impact parameter b the solar radius. Setting X equal to the Earth–Sun distance, then for return pulses grazing the Sun from a planet at distance, say $2X$ from Earth, the total time delay for the out-and-return journey will be $4\Delta t \sim 230\mu$s.

2.8 Orbital precession

Another early success of general relativity was the correct prediction of the tiny precession of the axes of the elliptical orbits of the planets. The historical case was that of the planet Mercury. The actual precession observed was of 532 arcsec per century, of which all but 43 arcsec could be accounted for by tidal forces due to other planets. Before Einstein's theory, the discrepancy had been attributed to the existence of an extra planet (Vulcan) which, however, was never found. According to the general theory, for the case of 'weak fields' there is, in addition to the usual $1/r^2$ Newtonian term, a small extra term varying as $1/r^4$ in the gravitational force. This has the result that the orbit is no longer a totally closed ellipse with fixed axes, but can be interpreted as a closed elliptic orbit in which the axis advances slowly with time. The resulting advance in the angle of rotation φ can be calculated, starting from the Schwarzschild line element (2.21), and applying the Euler–Lagrange equation (see Chapter 3) to deduce the equation of the orbit. For a circular orbit of radius r the fractional advance is found to be (for $r \gg r_s$):

$$\frac{\Delta\varphi}{\varphi} = \frac{3r_s}{2r} = \frac{3u^2}{c^2} \tag{2.30}$$

where $r_s = 2GM/c^2$ is the Schwarzschild radius of the Sun, mass M, and u is the orbital velocity (with $u^2 = GM/r$). Again we may note that this formula can be understood on the basis of the equivalence principle and special relativity, as follows. The angular momentum of the planet $\mathbf{L} = m(\mathbf{r} \times \mathbf{u}) = mr^2 \mathrm{d}\varphi/\mathrm{d}t$ is a constant[1] (Kepler's second law). Using the same arguments following equation (2.24), and transforming from the accelerated frame to an IF, the quantity r^2 will be contracted by a factor $(1 - r_s/r)$, while the time t will be dilated by a factor $1/(1 - (r_s/2r))$. So φ (or $\int \mathrm{d}t/r^2$) must be increased by a factor $(1 + 3r_s/2r)$.

[1] $\mathbf{L}$ is normal to the orbit plane, so is unaffected by the above transformation.

For an elliptic orbit of eccentricity e and semimajor axis a, the radius vector r is given by

$$\frac{1}{r} = \frac{(1 + e\cos\varphi)}{\left[a\left(1 - e^2\right)\right]}$$

Integrating over φ to find an averaged value, $\int (1/r)\,\mathrm{d}\varphi/2\pi$, it follows that for an elliptic orbit, $1/r$ in (2.30) must be replaced by $1/\left[a\left(1 - e^2\right)\right]$. From the values of the constants for the Sun and Mercury (see Problem 2.1), the above formula yields a value for the precession of 42.9 arcsec per century. The fact that the calculated and observed precessions agreed within 1% was a dual triumph, for the theory of relativity and for the precision measurements of the astronomers. Larger orbital precessions are expected for stronger fields, the most marked example so far being the 17° precession per year for the twin pulsars PSR J0737-3039 (see Chapter 10).

It should be noted that all the tests in Sections 2.6–2.8 are for *weak gravitational fields*, providing small, linear, perturbative corrections to Newtonian mechanics. General relativity has emerged from these and other tests with flying colours. However, more crucial tests would be for very strong fields, where the effects of non-linearities in the Einstein field equations should be manifest.

2.9 The Robertson–Walker line element

The Friedmann–Lemaitre–Robertson–Walker model of the universe (FLRW for short), which forms part of the 'Standard Model' of cosmology, is discussed in Chapter 5. Here, we just note that this is the simplest model, based on the concept of an originally isotropic and homogeneous expanding universe with a uniform space–time curvature. The line element (2.9) appropriate to IFs in special relativity is modified in general relativity for the FLRW model as follows:

$$\mathrm{d}s^2 = c^2\mathrm{d}t^2 - R(t)^2\left[\frac{\mathrm{d}r^2}{\left(1 - kr^2\right)} + r^2\,\mathrm{d}\theta^2 + r^2\sin^2\theta\,\mathrm{d}\phi^2\right] \tag{2.31}$$

where $R(t)$ is a universal expansion parameter, which multiplies the radial space coordinate r defined in a reference frame co-moving (i.e. enlarging) with the expansion, so that the physical coordinate distance D from one point to another (anywhere in the universe since it is assumed to be isotropic and homogeneous) is written as the product

$$D(t) = r \cdot R(t) \tag{2.32}$$

All the time dependence of distances between points is contained in the expansion factor $R(t)$. The time at which one measures R can be universal in this model, since observers can in principle be scattered all over the universe, and they can agree to synchronize their identical clocks when the universal density of expanding matter reaches a specified value. The parameter k describes the curvature of space. $k = +1$ corresponds to positive curvature, $k = -1$ to negative curvature and $k = 0$ to the flat Euclidean space of special relativity.

Equation (2.31) follows from general relativity for the particular case of an isotropic and homogeneous universe undergoing an isotropic expansion with a uniform curvature k/R^2, and we have simply stated the result. However, the form of the curvature term affecting the radial coordinate can be understood from a two-dimensional analogy. Figure 2.8 shows a section through a sphere of radius ρ, centre O, and two points on the surface of the sphere A and B. The shortest distance between A and B along the surface is the arc AB of a great circle, of length $2l$. Denote the angle subtended at the centre by $2\alpha = 2l/\rho$ and the chord AB through the sphere by $2D$, where $D = \rho\sin\alpha$. Then

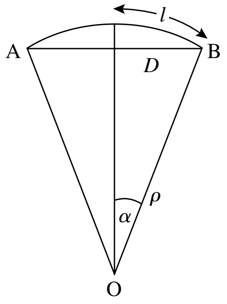


Fig. 2.8

$$D = \rho \sin\left(\frac{l}{\rho}\right)$$

$$\mathrm{d}D = \cos\left(\frac{l}{\rho}\right)\,\mathrm{d}l$$

and

$$\mathrm{d}l = \frac{\mathrm{d}D}{\cos(l/\rho)} = \frac{\mathrm{d}D}{\sqrt{1 - D^2/\rho^2}}$$

If we define the curvature $1/\rho^2 = k/R^2$, where $k = +1$ in this case, then with $D = Rr$ as above, we obtain

$$\mathrm{d}l = \frac{R\,\mathrm{d}r}{\sqrt{1 - kr^2}} \tag{2.33}$$

This is the form in (2.31) expressing the element of arc length dl along the surface of the sphere in terms of the curvature parameter k and the element of chord length $\mathrm{d}D = R\mathrm{d}r$. This two-dimensional analogy in fact carries straight over to three (or any larger number) of spatial dimensions, because of the spatial isotropy assumed in the model. However, this analogy applies for $k = +1$ only and there is no two-dimensional analogue for $k = -1$.

When the FLRW metric (2.31) is inserted into the field equations (2.19), there results the very important Friedmann equation(s) described in Chapter 5. Equation (2.31) is one of the key equations in cosmology and will be used frequently in the following chapters.

2.10 Modifications to Newtonian gravity?

The inverse square law of gravitational force (for weak gravitational fields and in the non-relativistic approximation) is known to hold with great accuracy for the solar system, that is, for distances within a few orders of magnitude of the astronomical unit. However, at much larger distances, that is, on the scale of our universe (giga parsec or 10^{26} m) it has been proposed that modifications may occur, and that these are responsible for the fact that 95% of the energy density of the universe appears to be due to the existence of completely new forms of dark matter/dark energy, as described in Chapter 7. At the present time, despite extensive searches, there is no *direct* experimental evidence for dark matter, for example, in the form of new elementary particles which can be detected and measured in the laboratory. So, could it be that it is all an artefact of deviations from the inverse square law at very large distances? As described in Section 7.4, there is absolutely no indication that this is the case, and that indeed examples are known of distant colliding galaxies in which the visible, luminous matter, subject to electromagnetic forces in collision, is well separated from the dark matter, subject to and identified from its gravitational (lensing) effects. No modifications to the inverse square law could possibly account for these phenomena.

One can also question the inverse square law at very small distances. The energy density of dark energy (about 5 GeV m^{-3}) when expressed in natural units ($\hbar = c = 1$), can be used to define a fundamental length, of about 85 μm (see Section 7.4). Could it be that the dark energy is somehow associated with deviations from the inverse square law at such distances, connected with the idea that the weakness of gravity compared with the other interactions is due to extra dimensions which at normal energies, well below the Planck scale, are curled up and ineffective? The torsion balance experiments of Kapner *et al.* (2007), however, have verified Newton's law with great accuracy down to distances of around 50 μm.

To summarize, over the years physicists, for one reason or another, have questioned the validity of Newton's inverse square law (always of course in the static, non-relativistic approximation to the Einstein equations) over distance scales covering at least 30 orders of magnitude. All the evidence so far, however, corroborates Newton's law.

2.11 Relativistic kinematics: four-momentum; the Doppler effect

The transformations of energy and momentum of a particle between IFs in special relativity are easily found from the coordinate transformations (2.1) by replacing x, y, z by the Cartesian components p_x, p_y, p_z of the three-momentum **p** of the particle, and the time component t by the total energy E (using units $c = 1$ for brevity). Then the transformations between an IF Σ and another frame Σ' moving at velocity βc in the x-direction with respect to Σ are

$$\begin{aligned} p'_x &= \gamma\left(p_x - \beta E\right) \\ p'_y &= p_y \\ p'_z &= p_z \\ E' &= \gamma\left(E - \beta p_x\right) \end{aligned} \tag{2.34}$$

where $\gamma = (1 - \beta^2)^{-1/2}$, and the quantity

$$E'^2 - p'^2 = E^2 - p^2 = m^2 \tag{2.35}$$

where m, the rest-mass of the particle, is invariant under the transformation.

The above transformations give the formula for the Doppler effect in relativistic optics. Consider a light beam of frequency ν and quantum energy $E = h\nu$ emitted in the frame Σ at an angle θ with respect to the x-axis. From (2.34) the energy $E' = h\nu'$ in the frame Σ' will be (since $p = E$ for photons in our units of $c = 1$)

$$E' = \gamma E\,(1 - \beta\cos\theta)$$

and for $\theta = 0$,

$$\nu' = \gamma\nu(1 - \beta) = \nu\sqrt{\frac{(1-\beta)}{(1+\beta)}} \tag{2.36}$$

Thus a source of light travelling at velocity βc away from an observer has its frequency 'red-shifted' by a factor ν'/ν. For $\theta = \pi/2$, one obtains a transverse Doppler effect, with

$$\nu' = \gamma\nu = \nu\left(1 + \frac{\beta^2}{2} + \cdots\right),$$

so it is second-order in β for small values of β, while for $\theta = 0$ it is first order.

We already saw in (2.8) that the expression for the square of the interval ds was an invariant under a transformation to another IF. The quantity ds is called a *four-vector*, since it has four space and time components, and the invariance under Lorentz transformations is analogous to the invariance in three space dimensions, of the length squared of a three-component vector under translations or rotations of the coordinate axes.

In kinematics we express the quantities in (2.34) as components of four-vectors called *four-momenta* p_μ ($\mu = 0, 1, 2, 3$), where for a single particle $p_0 = E, p_1 = p_x, p_2 = p_y, p_3 = p_z$. The four-momentum squared is a Lorentz scalar with the value (in units $c = 1$)

$$p^2 = E^2 - |\mathbf{p}|^2 = m^2 \tag{2.37}$$

In scattering experiments in high-energy physics, the result of the scattering of a particle by the interaction can be expressed in terms of the invariant *four-momentum transfer* $q_\mu = p_\mu - p'_\mu$, so that $q^2 = (E - E')^2 - (\mathbf{p} - \mathbf{p}')^2$, where the unprimed and primed quantities refer to the values for a particle before and after the interaction. It is left as an exercise to show that q^2 is always negative in a scattering process.

In kinematic problems, it is advantageous to evaluate quantities in the centre-of-momentum system (CMS) that is in a coordinate frame where the total three-momentum $\mathbf{p}$ of the colliding particles is zero. Then the invariant four-momentum squared (which can be evaluated in any IF) is just equal to the square of the total CMS energy, conventionally denoted by the symbol $s = E^2_{CMS}$ (but not to be confused with the space–time interval, also called s).

2.12 Fixed-target and colliding-beam accelerators

As an example of the use of four-vectors and the centre-of-momentum frame of reference, we consider fixed-target and colliding-beam accelerators. Suppose a beam of particles of mass m_a, energy E_a, and three-momentum $\mathbf{p}_\mathrm{a}$collides with a target particle of mass m_{b}, energy $E_{\mathrm{b},}$ and momentum $\mathbf{p}_{\mathrm{b}}$. Then the total four-momentum squared is given by

$$s = (E_{\mathrm{a}} + E_{\mathrm{b}})^2 - \left(\mathbf{p}_{\mathrm{a}} + \mathbf{p}_{\mathrm{b}}\right)^2 = m_{\mathrm{a}}^2 + m_{\mathrm{b}}^2 + 2\left(E_{\mathrm{a}}E_{\mathrm{b}} - \mathbf{p}_{\mathrm{a}} \cdot \mathbf{p}_{\mathrm{b}}\right) \tag{2.38}$$

The energy available for new particle creation is $\varepsilon = (\sqrt{s} - m_{\mathrm{b}} \text{ - } m_{\mathrm{a}})$. If $E_{\mathrm{a}} \gg m_{\mathrm{a}}$ and $E_{\mathrm{b}} \gg m_{\mathrm{b}}$, then

$$\varepsilon^2 \approx 2\left(E_{\mathrm{a}}E_{\mathrm{b}} - \mathbf{p}_{\mathrm{a}} \cdot \mathbf{p}_{\mathrm{b}}\right) \tag{2.39}$$

(a) Fixed Target

If the beam of *a* particles collides with a stationary target *b*, so that $E_{\mathrm{b}} = m_{\mathrm{b}}$ and $\mathbf{p}_{\mathrm{b}} = 0$, then

$$\varepsilon \approx (2m_{\mathrm{b}}E_{\mathrm{a}})^{1/2} \tag{2.40}$$

and the available energy rises with the square root of the incident energy. Examples of accelerators that have used fixed targets are the CERN PS (Proton Synchrotron) accelerating protons to 28 GeV, and the CERN SPS (Super Proton Synchrotron) for protons up to 400 GeV.

(b) Colliding beams

If the beams of *a* and *b* particles collide head-on, then $\mathbf{p}_{\mathrm{a}} \cdot \mathbf{p}_{\mathrm{b}} = -|\mathbf{p}_{\mathrm{a}}||\mathbf{p}_{\mathrm{b}}|$ and assuming both beams are extreme relativistic, we obtain

$$\varepsilon \approx \left[2\left(E_{\mathrm{a}}E_{\mathrm{b}} + |\mathbf{p}_{\mathrm{a}}|\,|\mathbf{p}_{\mathrm{b}}|\right)\right]^{1/2} \approx \sqrt{4E_{\mathrm{a}}E_{\mathrm{b}}}$$

In many colliders, the two beams are of particles of equal masses and equal energies, that is, $E_{\mathrm{a}} = E_{\mathrm{b}} = E$, when

$$\varepsilon \approx 2E \tag{2.41}$$

so that the available energy rises in proportion to the beam energy. An example is the LEP II $\mathrm{e}^+\mathrm{e}^-$ collider at CERN, which accelerated electrons and positrons in opposite directions in the same vacuum tube. The beam energies were $E =$ 100 GeV, so that the 200 GeV available CMS energy was sufficient to investigate the reactions $\mathrm{e}^+\mathrm{e}^- \rightarrow W^+W^-$ and $Z^0\bar{Z}^0$, where the threshold energies are $2M_W \sim 160$ GeV and $2M_Z \sim 180$ GeV. The HERA machine is an example of an asymmetric ep collider, accelerating electrons or positrons to 28 GeV energy in one vacuum ring, and protons to 820 GeV in the other direction in a second ring above the first, the two beams being brought into collision in two intersection regions. In this case, the square of the CMS energy is $s = 93{,}000\ \mathrm{GeV}^2$, and the useful maximum value of four-momentum transfer squared between the particles is $|q^2(\mathrm{max})| \sim 20{,}000\ \mathrm{GeV}^2$, about 100 times the maximum useful value attainable using secondary muon or neutrino beams from fixed-target machines.

Colliding-beam machines have the obvious advantage of providing much higher CMS energies for a given beam energy, and have been essential in identifying the more massive fundamental particles—the $W^{\pm}$ and Z^0 mediators of the electroweak interactions and the bottom and top quarks during the 1980s and 1990s. They are, however, limited to beams of stable or nearly stable particles, namely electrons, positrons, protons, antiprotons, heavy ions, and possibly in the future, muons (in $\mu^+\mu^-$ colliders).

Fixed-target machines achieve lower CMS energies but have the advantage that they can produce a range of intense high-energy beams of secondary particles, for example, π, $K^\pm$, K^0, μ, ν_μ. These were important historically, in laying the quantitative experimental foundations of particle physics, including the establishment of the quark substructure of matter and of CP violation in weak interactions in the 1960s, and of the electroweak theory and quantum chromodynamics of the strong interquark interactions in the 1970s. They still have very important applications today both as injectors for colliders and as sources of neutrino beams for the study of neutrino oscillations.

Problems

At the end of the book, answers are given to all the problems. Full solutions are given to the more challenging problems, which are denoted by an asterisk.

(2.1) Calculate the precession of the orbit of Mercury from (2.30), given the following data:

$$\begin{aligned} \text{Mercury mass} &= 3.24 \times 10^{23}\,\text{kg.} \\ \text{Semimajor axis} &= 0.387\,\text{A.U.} \\ \text{Eccentricity} &= 0.206 \\ \text{Earth mass} &= 5.98 \times 10^{24}\,\text{kg.} \\ \text{Sun mass} &= 1.99 \times 10^{30}\,\text{kg.} \end{aligned}$$

(2.2) Calculate the maximum value of the square of the momentum transfer q^2 in (GeV/c)^2, in the head-on collision at the HERA collider of a 28 GeV electron with a quark carrying 20% of the energy of an 800 GeV proton.

(2.3) The neutral pion undergoes the decay $\pi^0 \rightarrow 2\gamma$. Because the pion has spin zero, the angular distribution of the γ-rays in the pion rest frame is isotropic. The γ-rays from decay of pions produced by cosmic ray collisions in the atmosphere are observed. Show that their energy spectrum has a peak intensity for $E_\gamma = m_\pi c^2/2$, and that if E_1 and E_2 are the two γ energies on either side of this maximum for which the γ-ray intensities are equal, then $\sqrt{E_1E_2} = m_\pi c^2/2$. (*Note*: this method was used in 1950 to obtain one of the earliest measurements of the neutral pion mass).

(2.4) A photon of wavelength λ is emitted from the solar surface. Calculate the shift $\Delta\lambda$ in wavelength of the photon when it arrives at the Earth's surface.

(Solar mass and radius 1.99×10^{30} kg, 6.96×10^8 m

Earth mass and radius 5.98×10^{24} kg, 6.37×10^6 m.)

*(2.5) Using the transformations of momentum and energy in Section (2.11), derive an expression for the laboratory angle of emission of a γ-ray by an unstable particle in terms of the angle of emission in the rest-frame of the particle. Thus show that for a beam of very high-energy particles, half the γ-rays will be concentrated inside a forward cone of opening angle θ of order $1/\gamma$.

(2.6) A satellite travels in a circular orbit about the Earth with a period of 12 h. Calculate the fractional difference (with sign) in time between a satellite clock and an identical clock on the Earth. Refer to Question (2.4) for the Earth mass and radius. Neglect the effect of the Earth's rotation. (This problem has a practical application in correcting the times on the atomic clocks of the satellites used in the Global Positioning System.)

3 Conservation rules, symmetries, and the Standard Model of particle physics

3.1 Transformations and the Euler–Lagrange equation

One of the most important concepts in physics is that of the symmetry or invariance of a system under a particular operation. For example, a snowflake is invariant under a 60° rotation in the plane of the flake, and this tells us something about the physics of the molecular bonding in water. In fact, conservation rules and the associated symmetries have been called the backbone of high-energy particle physics.

Not all conservation rules are absolute. While conservation of electric charge or of energy and momentum are, as far as we know, sacrosanct in all situations, some quantities—for example, conservation of parity—that is, invariance under spatial reflection—hold for certain types of fundamental interaction but not for others. Moreover, on the enormous distance and time scales of the universe at large, it turns out that some quantum numbers, such as baryon number, which seem to be exactly conserved under laboratory conditions, are violated in cosmology. This violation presumably occurred at a very early, very hot stage of the universe, involving some new type of interaction at energies far above what can be obtained by accelerators on Earth.

The invariant description of physical phenomena in relativity, that is, under transformations of space and time coordinates, has been discussed in Chapter 2. In this chapter, we first of all discuss some of the important conservation rules and symmetries, and later describe the broken symmetries, including those which we believe have been at the core of the early development of the universe, described also in Chapter 4.

One familiar conservation law is that of linear momentum in classical mechanics, which follows if the energy of the system is invariant under translations in space. For, if there is no change in energy under such a translation, there can be no external forces on the system and the rate of change of momentum must be zero. This last example can be formalized by the Euler–Lagrange equation of classical mechanics. This is based on Hamilton's 'principle of least action'. The Lagrangian function in this case is the difference of kinetic and potential energies, $L = T - V$, for a set of particles. The action $S = \int L\mathrm{d}t$ and the above principle states that the path between fixed starting and ending coordinates $q(t_1)$ and $q(t_2)$ travelled by a particle between times t_1

and t_2 is such that the action is at an extremum (corresponding in fact to the shortest path in space or to the maximum path in proper time). For example, denoting the space coordinate by q and the velocity $\dot{q} = \mathrm{d}q/\mathrm{d}t$ as independent variables, we can write for the perturbation in the action due to a variation δq from the 'classical path' traced by the particle as

$$\delta S = \int \left[\left(\frac{\partial L}{\partial q} \right) \delta q + \left(\frac{\partial L}{\partial \dot{q}} \right) \delta \dot{q} \right] \mathrm{d}t = 0$$

Since $\delta \dot{q} = \mathrm{d}\left(\delta q\right)/\mathrm{d}t$, we can integrate the second term in the integrand by parts to obtain

$$\int \left(\frac{\partial L}{\partial \dot{q}} \right) \delta \dot{q}\, \mathrm{d}t = \left[\delta q \left(\frac{\partial L}{\partial \dot{q}} \right) \right] - \int \delta q \left[\frac{\mathrm{d}\left(\partial L/\partial \dot{q}\right)}{\mathrm{d}t} \right] \mathrm{d}t$$

For the limits of integration, which are at the fixed endpoints of the path, $\delta q = 0$, so the first term on the right-hand side vanishes. Hence,

$$\partial S = \int \left\{ \left(\frac{\partial L}{\partial q} \right) - \frac{\mathrm{d}\left(\partial L/\partial \dot{q}\right)}{\mathrm{d}t} \right\} \delta q \;\mathrm{d}t = 0$$

Since this must be true for arbitrary values of δq, the integrand must be zero, and the Euler–Lagrange equation follows:

$$\frac{\partial L}{\partial q} - \frac{\mathrm{d}}{\mathrm{d}t} \left(\frac{\partial L}{\partial \dot{q}} \right) = 0 \tag{3.1}$$

Thus if L is independent of q, $\partial L/\partial q = 0$ and the momentum $p = \partial L/\partial \dot{q}$ is constant. Here a global symmetry—invariance of L under space translations—has led to a conservation law. In this case we have taken the independent variables as the space coordinate q and the momentum p, but (3.1) can represent any pair of generalized coordinates, one being a derivative of the other.

In relativistic quantum mechanics, the Lagrangian function L is a field *energy density* rather than a sum over the energies of discrete particles, and it is furthermore a function of both space and time. The global invariance of L under space–time translations leads to a *conserved current* of four momentum, which is an example of a more general theorem called Noether's theorem, discussed in Section 3.8 below. The conservation of the fourth (time) component of this current corresponds to conservation of energy, and of the three space components to conservation of momentum.

The invariance (or non-invariance) of a physical system may occur for *continuous* transformations, for example, a rotation in a phase angle or a translation in space; or it can be a *discrete* transformation, such as the inversion of a spatial or time coordinate, or charge conjugation. For continuous transformations, the associated conservation laws and quantum numbers are additive (the total conserved energy of system is equal to the sum of the energies of its parts), while for discrete transformations they are multiplicative (e.g. the symmetry under spatial reflection, called the parity, is equal to the product of the parities of the parts of the system).

3.2 Rotations

As an example of a continuous transformation let us consider a spatial rotation through some angle, say ϕ, about the z-axis. The operator of the z component of angular momentum in Cartesian coordinates is defined as

$$J_z = -i\hbar \left(x\frac{\partial}{\partial y} - y\frac{\partial}{\partial x} \right)$$

This operation can also be described by a rotation. Suppose a vector of length r lying in the xy plane makes an angle ϕ with the x-axis. Then in a rotation through $\delta\phi$, the increments of the Cartesian components are

$$\delta y = r\cos\phi \; \delta\phi = x\delta\phi$$
$$\delta x = -r\sin\phi \; \delta\phi = -y\delta\phi$$

The effect of this rotation on a function $\psi(x, y, z)$ will be

$$R\,(\phi,\, \delta\phi)\;\; \psi\left(x, y, z\right) = \psi\left(x + \delta x, y + \delta y, z\right) = \psi\left(x, y, z\right) + \delta x\left(\frac{\partial\psi}{\partial x}\right)$$
$$+\,\delta y\left(\frac{\partial\psi}{\partial x}\right) = \psi\left[1 + \left(x\frac{\partial}{\partial y} - y\frac{\partial}{\partial x}\right)\right] = y\left[1 + \delta\phi\frac{\partial}{\partial\phi}\right]$$

Hence the J_z operator

$$J_z = -i\hbar\left[x\frac{\partial}{\partial y} - y\frac{\partial}{\partial x}\right] = -i\hbar\frac{\partial}{\partial\phi}$$

A finite rotation is achieved by making n infinitesimal steps, that is, $\Delta\phi = n\delta\phi$ where $n \to \infty$, so that

$$R = \operatorname*{Lim}_{n\to\infty}\left(1 + iJ_z\frac{\delta\phi}{\hbar}\right)^n = \exp\left(iJ_z\frac{\Delta\phi}{\hbar}\right) \qquad (3.2)$$

Here J_z is said to be the *generator* of the rotation $\Delta\phi$.

3.3 The parity operation

The inversion of spatial coordinates $(x, y, z) \to (-x, -y, -z)$ is a discrete transformation on the wave amplitude ψ brought about by the parity operator $P : P\psi(\mathbf{r}) = \psi(-\mathbf{r})$. Since in repeating the operation one reverts to the original system, $P^2 = 1$ and the eigenvalues of P must be ± 1. These eigenvalues are referred to as the *parity* of the system ψ. For example, the function $\psi = \cos x$ has $P = +1$ or positive parity since $P\psi = P(\cos x) = \cos(-x) = +\psi$, while if $\psi = \sin x, P\psi = -\psi$ so that ψ has negative parity. On the other hand, in the case of a function such as $\psi = (\sin x + \cos x), P\psi = (\cos x - \sin x) \neq \pm\psi$ so this function is not an eigenstate of parity. Parity is a useful concept when dealing with elementary particles since the interactions often have very well-defined properties under the parity operation. This may be contrasted with biological systems, for example, a runner bean or the DNA molecule, which are not eigenstates of parity.

A spherically symmetric potential has the property that $V(-\mathbf{r}) = V(\mathbf{r})$, so that states bound by such a potential—as is usually the case in atoms—can be parity eigenstates. For the hydrogen atom, the wavefunction in terms of the radial coordinate r and the polar and azimuthal angular coordinates θ and ϕ of the electron with respect to the proton is

$$\chi(r,\theta,\phi) = \eta(r)\; Y_l^m(\theta,\phi)$$

where Y is the spherical harmonic function, with l the orbital angular momentum quantum number and m its z-component. Under inversion, $\mathbf{r} \to -\mathbf{r}, \theta \to (\pi - \theta)$, while $\phi \to (\pi + \phi)$ with the result that

$$Y_l^m(\pi - \theta, \pi + \phi) = (-1)^l\; Y_l^m(\theta,\phi)$$

Hence in this case,

$$P\chi(r,\theta,\phi) = (-1)^l\, \chi(r,\theta,\phi) \tag{3.3}$$

3.4 Parity conservation and intrinsic parity

In strong and electromagnetic interactions, parity is found to be conserved: the parity in the final state of a reaction is equal to that in the initial state. For example, for an electric dipole ($E1$) transition in an atom, the change in l is governed by the selection rule $\Delta l = \pm 1$. Thus from (3.3) the parity of the atomic state must change in such transitions, which are accompanied by the emission of photons of negative parity, so that the parity of the whole system (atom + photon) is conserved. For a (less probable) magnetic dipole ($M1$) transition, or for an electric quadrupole ($E2$) transition, the selection rules are $\Delta l = 0$ and 2 respectively, and in either case the radiation is emitted in a positive parity state. In high-energy physics, one is generally dealing with pointlike or nearly pointlike interactions and electromagnetic transitions involving small changes in angular momentum ($\Delta J = \pm 1$), in which case photons are emitted with negative parity.

The symmetry of a pair of identical particles under interchange, which was described in Section 1.3, can be extended to include both spatial and spin functions of the particles. If the particles are non-relativistic, the overall wavefunction can be written as a simple product of space and spin functions:

$$\psi = \chi\,(\text{space})\,\alpha\,(\text{spin})$$

Consider two identical fermions, each of spin $s = 1/2$, described by a spin function $\alpha(S, S_z)$ where S is the total spin and $S_z = 0$ or ± 1 is its component along the z-(quantization) axis.

Using up and down arrows to denote z-components of $s_z = +1/2$ and $-1/2$, we can write down the $(2s + 1)^2 = 4$ possible states as follows:

$$\left.\begin{aligned} \alpha(1,+1) &= \uparrow\uparrow \\ \alpha(1,-1) &= \downarrow\downarrow \\ \alpha(1,0) &= (\uparrow\downarrow + \downarrow\uparrow)/\sqrt{2} \end{aligned}\right\} \quad S = 1,\ \text{symmetric}$$

$$\alpha(0,0) = (\uparrow\downarrow - \downarrow\uparrow)/\sqrt{2} \qquad S = 0, \text{antisymmetric} \tag{3.4}$$

The first three functions are seen to be symmetric under interchange, that is, α does not change sign, while for the fourth one it does. It is seen that the sign of the spin function under interchange is $(-1)^{S+1}$ while that for the space wavefunction from (3.3) is $(-1)^L$, where L is the total orbital angular momentum. Hence the overall sign change of the wavefunction under interchange of both space and spin coordinates of the two particles is

$$\psi \to (-1)^{L+S+1}\,\psi \tag{3.5}$$

As an example of the application of this rule, let us consider the determination of the intrinsic parity of the pion. This follows from the existence of the S-state capture of a negative pion in deuterium, with the emission of two neutrons:

$$\pi^- + d \to n + n \tag{3.6}$$

The deuteron has spin 1, the pion spin 0, so that in the initial state and therefore in the final state also, the total angular momentum must be $J = 1$. If the total spin of the neutrons is S and their orbital angular momentum is L, then $\mathbf{J} = \mathbf{L} + \mathbf{S}$. If $J = 1$ this allows $L = 0, S = 1$; or $L = 1, S = 0$ or 1; or $L = 2, S = 1$. Since the neutrons are identical particles it follows that their wavefunction ψ is antisymmetric, so that from (3.5) $L + S$ must be even and $L = S = 1$ is the only possibility. Thus the neutrons are in a ${}^3P_1$ state with parity $(-1)^L = -1$. The nucleon parities cancel on the two sides of (3.6), so that the pion must be assigned an *intrinsic parity* $P_\pi = -1$, so that parity be conserved in this strong interaction.

The assignation of an intrinsic parity to a particle follows if the particle can be created or destroyed *singly* in a parity-conserving interaction, in just the same way that electric charge has been assigned in the same interaction to obey charge conservation. Clearly, in the above reaction, the number of nucleons is conserved and so the nucleon parity itself is conventional. It is assigned $P_n = +1$. However, in an interaction it is possible, if the energy is sufficient, to create a nucleon–antinucleon pair, and hence determine its parity by experiment. So while the parity of a nucleon is fixed by convention, the *relative* parity of nucleon and antinucleon—or any other fermion–antifermion pair—is not.

In the Dirac theory of fermions, *particles and antiparticles have opposite intrinsic parity*. This prediction was verified in an experiment by Wu and Shaknov, shown in Fig. 3.1, using a ^{64}Cu positron source. Positrons from this source came to rest in the surrounding absorber and formed *positronium*, an 'atomic' bound state of electron and positron, which has energy levels akin to those of the hydrogen atom, but with half the spacing because of the factor 2 in the reduced mass. The ground level of positronium occurs in two closely spaced substates with different mean lifetimes: the spin-triplet (${}^3\mathrm{S}_1$) decaying to three photons (lifetime 1.4×10^{-7} s), and the spin-singlet state (${}^1\mathrm{S}_0$) decaying to two photons (lifetime 1.25×10^{-10} s). We consider here the singlet decay:

$$\mathrm{e}^+\mathrm{e}^- \to 2\gamma \tag{3.7}$$

The simplest wavefunctions describing the two-photon system, linear in the momentum vector $\mathbf{k}$ and in the polarization vectors ($\mathbf{E}$-vectors) $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$ of

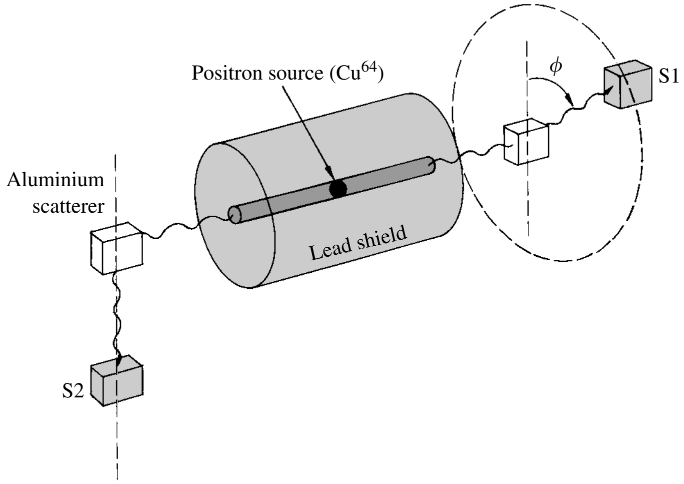


Fig. 3.1 Sketch of the method used by Wu and Shaknov (1950) to measure the relative orientation of the polarization vectors of the two photons emitted in the decay of $^1S_0$ positronium. $S1$ and $S2$ are two anthracene counters recording the gamma rays after scattering by aluminium cylinders. Their results proved that fermion and antifermion have opposite parity, as predicted by the Dirac theory of the electron.

the photons will be

$$\psi_1(2\gamma) = A\,(\varepsilon_1 \cdot \varepsilon_2) \qquad \propto \cos\phi \tag{3.8a}$$

$$\psi_2(2\gamma) = B\,(\varepsilon_1 \times \varepsilon_2) \cdot \mathbf{k} \qquad \propto \sin\phi \tag{3.8b}$$

where A and B are constants and ϕ is the angle between the planes of polarization. The first quantity ψ_1 is a scalar and therefore even under space inversion ($\phi \to -\phi$), thus requires positive parity for the positronium system. The quantity ψ_2 is the product of an axial vector with a polar vector, that is, a pseudoscalar quantity which is odd under inversion. It corresponds to a positronium system of negative parity, with a $\sin^2\phi$ distribution of the angle between the polarization vectors. In the experiment, the decays of singlet positronium were selected by observing the two photons emerging in opposite directions from a lead block. The photon polarization was determined indirectly by observing the Compton scattering off aluminium cubes, recorded in anthracene counters as shown in Fig. 3.1. The ratio of the scattering rates for $\phi = 90°$ and $\phi = 0°$ was 2.04 ± 0.08, consistent with the ratio of 2.00 expected for positronium of negative parity. Since the ground states of positronium are S-states, the parity measured is the same as that of the electron–positron pair. This experiment therefore confirms that fermions and antifermions have opposite intrinsic parity, as predicted by the Dirac theory.

3.5 Parity violation in weak interactions

While parity is conserved in the strong and electromagnetic interactions, it is violated—what is more, maximally violated—in the weak interactions. This is manifested in the observation that fermions participating in the weak interactions are *longitudinally polarized.* Let $\boldsymbol{\sigma}$ represent the spin vector of a particle of energy E, momentum $\mathbf{p}$, and velocity $\mathbf{v}$ travelling along the z-axis, with $\boldsymbol{\sigma}^2 = 1$. The longitudinal polarization P is the difference divided by the sum, of the numbers of particles N^+ and N^- with $\boldsymbol{\sigma}$ parallel and antiparallel to **p** (i.e. with spin components, in units of $h/2\pi$, of $\sigma_z = +1$ or -1) and is given by

$$P = \frac{(N^+ - N^-)}{(N^+ + N^-)} = \alpha\left(\boldsymbol{\sigma}\cdot\mathbf{p}\frac{c}{E}\right) = \alpha\frac{v}{c}$$
$$\begin{aligned} \text{where } \alpha &= -1 \text{ for fermions} \\ \alpha &= +1 \text{ for antifermions} \end{aligned} \tag{3.9}$$

This expression for the polarization of fermions in weak interactions was predicted in 1957 by the so-called V–A theory, applying to 'charged current' weak interactions, namely, those mediated by $W^\pm$ exchange. Figure 3.2 shows the experimental results on measurement of polarization of electrons emitted in nuclear beta decay, indicating $P = -v/c$ in support of the V–A theory.

Example 3.1 *Prove that a scalar meson ($J^P = 0^+$) cannot decay to three pseudoscalar mesons ($J^P = 0^-$) in a strong or electromagnetic interaction. Can it do so in a weak interaction?*

Let $\mathbf{k}_1$, $\mathbf{k}_2$, and $\mathbf{k}_3$ be the momenta of the three pseudoscalar mesons in the overall centre-of-momentum frame. Since all the particles are spinless, the decay amplitude can only be a function of their intrinic parities and their three momenta. The two possible linear combinations of the momentum vectors give the following expressions:

$\mathbf{k}_1 \cdot (\mathbf{k}_2 \times \mathbf{k}_3)$	pseudoscalar product
$\mathbf{k}_1 \cdot (\mathbf{k}_2 - \mathbf{k}_3)$	scalar product

Since the parent meson is scalar and the product particles have intrinsic parity $(-1)^3 = -1$, we need to take the pseudoscalar product. Since $\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3 = 0$, it follows that $\mathbf{k}_1 \cdot (\mathbf{k}_2 \times \mathbf{k}_3) = -\mathbf{k}_1 \cdot \mathbf{k}_2 \times (\mathbf{k}_1 + \mathbf{k}_2) =$

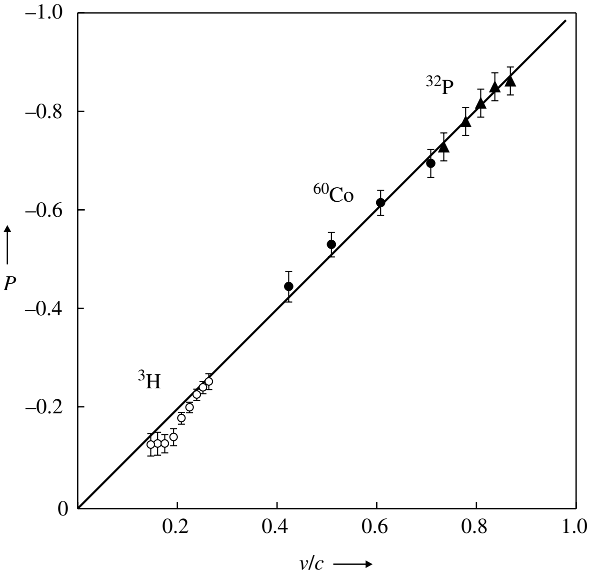


Fig. 3.2 The longitudinal polarization of electrons emitted in nuclear beta decay, plotted as a function of electron velocity v. (After Koks and Van Klinken (1976).)

$-\mathbf{k}_1 \cdot \mathbf{k}_2 \times \mathbf{k}_1 = 0$, since the three momentum vectors must be coplanar and the amplitude vanishes. If the decay is a weak process, parity is not conserved, so that the scalar product above can be involved and the amplitude can be finite.

3.6 Helicity and helicity conservation

For ultra-relativistic particles with $v = c, |pc| = E$, the polarization (3.9) has the simple form

$$H = \frac{\boldsymbol{\sigma} \cdot \mathbf{p}}{|p|} = +1 \text{ or } -1 \tag{3.10}$$

where the quantity H is called the *helicity* or handedness. Neutrinos have extremely small masses, that is, velocities $v \approx c$. The momentum and spin vectors define a screw sense, with neutrinos being left-handed (LH) and antineutrinos right-handed (RH)—see Fig. 3.3(a).

Neutrinos are eigenstates of helicity, with $H = -1$, while antineutrinos have $H = +1$. This is a relativistically invariant description: in transforming from the lab frame to another reference frame, necessarily with velocity $v < c$, it is impossible to change the sign of the helicity.

On the other hand, particles with finite mass such as electrons cannot exist in pure helicity eigenstates; they are mixtures of positive and negative helicity states. For example, electrons emitted in weak interactions (e.g. in nuclear beta decay) with velocity v are longitudinally polarized, consisting of a combination of LH states with intensity $1/2\,(1 + v/c)$ and RH states with intensity $1/2\,(1 - v/c)$, so that the net polarization $P = -v/c$ as in (3.9).

In the interactions of high-energy particles, there is a simple rule about helicity. For interactions involving *vector or axial–vector fields, helicity is conserved in the relativistic limit*. Note that the strong, electromagnetic, and weak interactions are all mediated by vector or axial–vector bosons (G, γ, W, or Z exchanges). This means that, in any such interaction, and provided the particle involved is relativistic, its helicity is preserved. Thus a high-energy electron of $v \approx c$ in a LH state, for example, will remain in a LH state as it emerges from the interaction. This helicity rule determines the angular distribution in many high-energy interactions, and is well illustrated in the interactions of neutrinos and electrons shown in Fig. 3.3(b–d).

The cross-sections given in the caption to Fig. 3.3 and in the Example 3.2 below are for high-energy scattering processes mediated by $W^{\pm}$ exchange—the so-called 'charged current' processes. In all these processes, 'neutral current' interactions mediated by Z^0 exchange will also contribute, and indeed this is the only possibility for $\mathrm{e}^+ + \mathrm{e}^- \rightarrow \nu_\mu + \bar{\nu}_\mu$ or $\nu_\tau + \bar{\nu}_\tau$. Such reactions will involve a weak mixing angle θ_w as described in Section 3.10. Summed over all flavours of neutrino and for both neutral and charged current reactions, the cross-section for electron–positron annihilation to neutrino–antineutrino pairs in the high-energy limit ($s \gg m_e^2$) is $\sigma \approx 1.3 G_\mathrm{F}^2 s/6\pi$. This annihilation process is of great astrophysical importance, both in the evolution of the early universe as described in Section 7.8 and in the later, supernova stages of giant stars, discussed in Section 10.8.

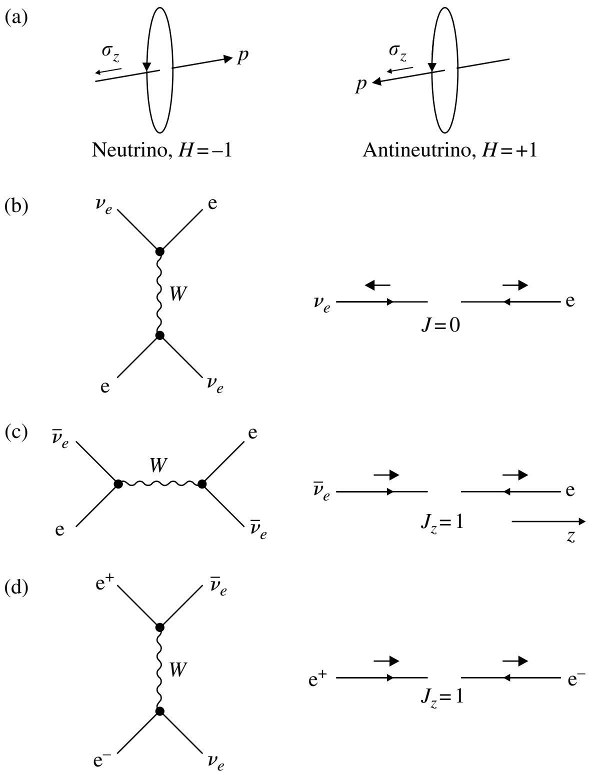


Fig. 3.3 (a) In the Standard Model, neutrinos are massless, and so are eigenstates of helicity. A neutrino has helicity $H = -1$, an antineutrino $H = +1$. Although neutrinos are not in fact massless, their masses, of order 0.1 eV/c^2, are so small that at normal laboratory energies, in the MeV–GeV range, they can be treated as effectively massless. The same is true for electrons, at energies large when compared with the electron mass. (b) The reaction $\nu_e + \mathrm{e} \to \nu_e + \mathrm{e}$ at high-energy mediated by $W^\pm$ exchange. Viewed in the CMS, both incident particles have $H = -1$ and the total angular momentum $J = 0$. Consequently, the angular distribution of the product particles is isotropic. The total cross-section, as given in (1.23), is $\sigma = G_F^2 s/\pi$, where s is the square of the CMS energy. (c) The diagram for antineutrino scattering, $\bar{\nu}_e + \mathrm{e}^- \to \bar{\nu}_e + \mathrm{e}^-$. As indicated in Example 3.2, the cross-section is one third of that in (b). (d) The diagram for $\mathrm{e}^+ + \mathrm{e}^- \to \nu_e + \bar{\nu}_e$. Relative to (c), the cross-section is reduced by a further factor of 2.

Example 3.2 *Calculate the cross-section for the process $e^+ + e^- \to \nu_e + \bar{\nu}_e$ via W-exchange, given that the cross-section for the process $\nu_e + e^- \to \nu_e + e^-$ is $G_F^2 s/\pi$. Assume the electron mass can be neglected at the energies involved.*

Let us do the calculation in two stages. First, we evaluate the cross-section for the scattering of antineutrinos by electrons. In the reaction $\bar{\nu}_e + e \to \bar{\nu}_e + e$, the incident electron and antineutrino have opposite helicities as in (3.9)—see Fig. 3.3(c). Hence $J = 1$ and $J_z = +1$. In this case, back-scattering of the antineutrino in the centre-of-momentum system (CMS) is forbidden by angular momentum conservation. Of the $2J + 1 = 3$ possible final states, only $J_z = +1$ is allowed by angular momentum conservation. Hence the cross-section, relative to that for neutrino scattering, is reduced by a factor 3 and $\sigma\,(\bar{\nu}_e + e \to \bar{\nu}_e + e) = G_F^2 s/3\pi$.

For the reaction $\mathrm{e}^+ + \mathrm{e}^- \to \nu_e + \bar{\nu}_e$, the incident leptons again have opposite helicities as in antineutrino–electron scattering. Just as in that reaction, only the LH helicity state of the electron can be coupled. The difference is that, while the antineutrino exists *only* in the RH state, the positron can have either LH or RH helicity (since it will have originated in an electromagnetic interaction). However, only the RH state of the positron can interact weakly with the (LH) electron, so that the cross-section is halved, and $\sigma\left(\mathrm{e}^+ + \mathrm{e}^- \to \nu_e + \bar{\nu}_e\right) = G_F^2 s/6\pi$.

3.7 Charge conjugation invariance

The operation of charge conjugation C reverses the sign of the electric charge and magnetic moment of a particle, leaving all other coordinates unchanged. Both strong and electromagnetic interactions are invariant under the C-operation. For example, Maxwell's equations are invariant under change of sign of the charge or current and thence of the fields **E** and **H**. In relativistic quantum mechanics, charge conjugation also implies particle–antiparticle conjugation, for example, $e^- \leftrightarrow e^+$. As an example of particle–antiparticle symmetry in electromagnetic interactions, a cyclic accelerator can accelerate electrons in a toroidal vacuum tube by means of radio-frequency cavities, and constrains them in, say, a clockwise circular path by means of a magnet ring. The *same* machine will equally accelerate positrons in an anti-clockwise direction, and this principle is used in electron–positron colliders, where the accelerated beams with equal energies are arranged, by means of bending and focusing magnets, to meet head-on once or many times per revolution.

On the contrary, weak interactions are not invariant under C. As shown in Fig. 3.3, a neutrino has $H = -1$, and the C-operation would transform it into an antineutrino of $H = -1$, which is a state that does not exist. However, the combined operation CP—charge conjugation followed by space inversion—would transform a LH neutrino into a RH antineutrino, which *does* exist (see Section 3.14 and Fig. 3.14 below).

Of course, to the extent that both lepton number (for the charged leptons) and baryon number are conserved, there can be no physical process turning an electron into a positron or a proton into an antiproton. However, neutral bosons, which are their own antiparticles, could be eigenstates of the C-operator. For example, under the C-operation the wavefunction of a neutral pion transforms into itself: $C\,|\pi^0\rangle \to \eta\,|\,\pi^0\rangle$ where, since repeating the process gets us back to the original state, $\eta^2 = 1$ and $C\,|\pi^0\rangle = \pm|\,\pi^0\rangle$. The neutral pion decays through an electromagnetic interaction, $\pi^0 \to 2\gamma$. The photon must have $C = -1$ since it is generated by charges and currents which reverse sign under the C-operation, and so for a system of n photons, $C = (-1)^n$. Thus the neutral pion must have $C = +1$, and the decay $\pi^0 \to 3\gamma$ is forbidden by C-invariance in electromagnetic interactions.

3.8 Gauge transformations and gauge invariance

In Section 3.1, we described examples of translations and rotations in physical space and time. What is just as significant for particle physics are the results of 'internal' symmetry transformations. For example, the plane wave function ψ representing a particle with four-momentum $p(= p_\mu$ where $\mu = 0, 1, 2, 3)$ can be modified by inserting an arbitrary phase factor α. If $x(= x_\mu)$ is the space–time coordinate then the transformation is (in units $\hbar = c = 1$):

$$\psi = \exp\left(ipx\right) \to \psi = \exp i\left(px + \alpha\right) \tag{3.11}$$

From (3.2) this operation is equivalent to a rotation in some internal 'charge space' of the particle. Clearly, if this phase transformation is *global* (i.e. the same over all space), it cannot affect any physical observable. For example, differentiating (3.11), one finds for the expectation value of the momentum of an electron

$$-\psi^* i \frac{\partial \psi}{\partial x} = p \tag{3.12}$$

where $-i\partial/\partial x$ is the momentum operator and the asterisk indicates complex conjugation. The result is independent of the choice of α, since the phase factors cancel. As indicated in Section 3.1, the invariance of the Lagrangian density under such a global phase transformation actually leads to a conserved current, *via* Noether's theorem. We can illustrate this by writing the above transformation as a small increment ($\alpha \ll 1$):

$$\psi \rightarrow \psi\,(1 + i\alpha) \tag{3.13}$$

The Lagrangian energy density L of the field ψ appears in the Euler–Lagrange equation analogous to the classical equation (3.1):

$$\frac{\partial}{\partial x}\left(\frac{\partial L}{\partial \psi'}\right) - \frac{\partial L}{\partial \psi} = 0 \tag{3.14}$$

where $\psi' = \partial\psi/\partial x$. If L is invariant under the transformation (3.13), then

$$\delta L = 0 = i\alpha\psi\left(\frac{\partial L}{\partial \psi}\right) + i\alpha\psi'\left(\frac{\partial L}{\partial \psi'}\right)$$

and since

$$i\alpha\frac{\partial}{\partial x}\left(\psi\frac{\partial L}{\partial \psi'}\right) = i\alpha\psi'\frac{\partial L}{\partial \psi'} + i\alpha\psi\frac{\partial}{\partial x}\left(\frac{\partial L}{\partial \psi'}\right)$$

then

$$\delta L = 0 = i\alpha\psi\left(\frac{\partial L}{\partial \psi} - \frac{\partial}{\partial x}\left(\frac{\partial L}{\partial \psi'}\right)\right) + i\alpha\frac{\partial}{\partial x}\left(\psi\frac{\partial L}{\partial \psi'}\right)$$

The first term on the right-hand side vanishes, from (3.14), so that the second term must also be zero. Thus if we denote the four-current by

$$J\,(= J_\mu) = \psi\left(\frac{\partial L}{\partial \psi'}\right),$$

this is conserved:

$$\frac{\partial J}{\partial x} = 0 \tag{3.15}$$

From (3.12) and the example of the Lagrangian for a scalar field in (3.25), we can see that this four-current has the dimensions of a four momentum. If we had included the electric charge $|e|$ as a factor in the phase transformation, then the above equation would represent conservation of electric current. Notice that in classical mechanics, invariance of the Lagrangian under some operation (e.g. translation in space) leads to a constant of the motion (in this case, the conserved three momentum), while in quantum mechanics, invariance of the Lagrangian density—a function of both space and time—under a global phase transformation leads to a conserved current.

So much for *global* phase transformations. However, it is also possible to make a *local* phase transformation, that is, one for which $\alpha = \alpha(x)$ is a function of the space/time coordinate. Then, including a factor $|e|$ in the phase to emphasize that we are dealing with electric currents

$$\frac{\partial\psi}{\partial x} = i\left(p + e\frac{\partial\alpha}{\partial x}\right)\psi$$

In this case, physically observable quantities like momentum *will* be affected by the choice of α and its x-dependence, so local phase invariance does not appear to be a useful concept. However, the electron is a charged particle and will therefore be subject to any electromagnetic potential, which will comprise a vector potential $\mathbf{A}$ and a scalar potential Φ. We know that the effect of Φ is to change the energy of the particle from E to $E - e\Phi$, and correspondingly, the four-vector potential $A = (\mathbf{A}, \Phi)$ will change the four-momentum from p to $(p - eA)$. Hence, if one includes the effects of an electromagnetic potential of arbitrary magnitude, the above derivative becomes

$$\frac{\partial\psi}{\partial x} = i\left(p - eA + e\frac{\partial\alpha}{\partial x}\right)\psi \tag{3.16}$$

The scale or gauge of the potential A is also arbitrary: one can add to it the gradient of any scalar function, without affecting the values of any physically measurable quantities, namely, the associated electric and magnetic fields. This change of the scale or gauge of the potential is called a *gauge transformation*. Choosing α as this arbitrary scalar function, the transformation $A \to A + \partial\alpha/\partial x$ gives for the derivative

$$\frac{\partial\psi}{\partial x} \to i\left(p - eA - e\frac{\partial\alpha}{\partial x} + e\frac{\partial\alpha}{\partial x}\right)\psi = i\left(p - eA\right)\psi \tag{3.17}$$

so that an observable quantity such as $\psi^*\partial\psi/\partial x$ no longer contains α or $\partial\alpha/\partial x$. The quantity $\partial/\partial x$ on the left-hand side has thus been replaced by $i(p - eA)$ or, in operator notation

$$\frac{\partial}{\partial x} \to D = \frac{\partial}{\partial x} - ieA \tag{3.18}$$

called the *covariant derivative*. Note that x and A here are four-vector quantities, that is the space–time coordinate $x = x_\mu(\mu = 0, 1, 2, 3$ with $x_0 = ct$, $x_1 = x$, $x_2 = y$, $x_3 = z)$, and similarly for the four-vector potential $A = A_\mu$, so that when written with the indices (3.18) becomes

$$D = \frac{\partial}{\partial x_\mu} - ieA_\mu \tag{3.19}$$

In summary, by judicious choice of the scalar function the effects of the original local phase transformation on the electron wavefunction and the gauge transformation on the potential cancel exactly. The fact that it is possible to formulate the theoretical description to have this property of local gauge invariance turns out to be vital for the quantum field theory of electromagnetism, called quantum electrodynamics (QED).

Intuitively, one can see on a qualitative basis that these global and local invariances must be consistent with *charge conservation* and the *masslessness* *of the photon*, respectively. Charge conservation on a global basis comes in because, if the electron were suddenly to lose and then later regain its charge, the above cancellation would not be perfect, since at some value of x the potential A would have no charge to operate on. So charge must be conserved globally. Second, since the electron can be located anywhere with respect to the source of potential, the electromagnetic field involved in the local gauge transformation must have indefinitely long range. From the discussion in Section 1.5 connecting the range of the interaction with the mass of the mediating boson, it follows that if the electromagnetic field has infinite range, the photon must be massless. A corollary is that mass terms cannot occur in the Lagrangian if there is gauge symmetry. As we shall see below, the electroweak theory *does* include massive bosons, and a special mechanism (the Higgs mechanism) is required to overcome this problem.

We may note also here that a *truly* massless photon is an idealized concept. Real photons have to originate somewhere and end up somewhere else, but the distance they can travel cannot exceed the optical horizon, which is the nominal radius of the observed universe, of order 10^{26} m. If we set this equal to the Compton wavelength λ of the photon, the limit on the mass would be $m_\gamma c^2 < \hbar c/\lambda \sim 10^{-32}$ eV. The best *experimental* limit on the photon mass is based on assuming equilibrium between the magnetic and gravitational fields in the Small Magellanic Cloud, giving a range exceeding 3 kpc (10^{20} m) and hence $m_\gamma c^2 < 10^{-27}$ eV.

Why do we stress the concept of gauge invariance? The point of a gauge-invariant theory is that it introduces a symmetry in the calculations, which makes the theory *renormalizable.* This means that it is possible, at least in principle, to make calculations in the form of a perturbation series to all orders in the coupling constant, that is, for a sum over all possible Feynman diagrams, including those involving an arbitrary number of exchanged photons, and not just the one photon exchange shown in Fig. 1.3.

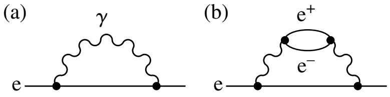


Fig. 3.4 (a) An electron is temporarily dissociated into an electron plus a virtual photon and (b) into an electron and a virtual electron–positron pair.

Figure 3.4 gives examples of how, in (a) an electron can be temporarily dissociated into a ‘bare’ electron of mass and charge m_0 and e_0, plus a virtual photon, and in (b) into an electron and a photon which converts to an e^+e^- pair. The first diagram involves the coupling α, the second α^2. Because in the second diagram the electric charges of the pair can affect and indeed reduce the field of the parent charge, this last process is referred to as ‘vacuum polarization’. Classically of course the vacuum contains nothing, by definition. In quantum mechanics, the definition is different: vacuum is defined as the state of lowest energy of the system. The uncertainty principle allows ‘vacuum fluctuations’, with an energy ΔE (in this example appearing in the form of a pair) provided these are limited to a time $\Delta t \sim \hbar/\Delta E$. Such fluctuations are further discussed in Chapter 7 in connection with dark energy in the universe; in Chapter 8 in the context of the inflationary model of the early universe, where they are postulated to account for the tiny anisotropies in the cosmic microwave background radiation; and in Chapter 10 in connection with Hawking radiation from black holes.

In calculating the effects of these so-called radiative corrections, a problem arises in that, in principle, the momentum k of the virtual particles involved can go to infinity, and their contribution to the energy of the system, which turns out to be of order $\int dk/k$, is therefore logarithmically divergent. If the mass of the ‘bare’ electron is denoted by m_0, this means that when the virtual

Table 3.1 Anomalous magnetic moments of electron and muon $(g-2)/2 \times 10^{10}$

	Predicted	Observed
Electron	$11,596,524 \pm 4$	$11,596,521.9 \pm 0.1$
Muon	$11,659,180 \pm 100$	$11,659,230 \pm 80$

particles are included, the value of m_0 will become infinite. In fact this idea of a bare mass is meaningless, since what the experimentalist actually measures is the electron *plus* all the associated virtual processes which can conceivably occur. In fact the same divergences are present in all the processes which the theorist calculates, and can be avoided in all the diagrams and to all orders in the coupling, by re-calibrating or *renormalizing* the (unobservable) bare charges and masses e_0 and m_0, to be their physically measured values, e and m, to be determined of course by experiment.

The correctness of the answers supplied by QED is illustrated by Table 3.1. The Dirac theory of a point lepton of mass m and charge e predicts a magnetic moment of 1 Bohr magneton, $\mu_B = e\hbar/2mc$. The actual moment is given by $\mu = \mu_B gs$ where $s = 1/2$ is the lepton spin in units of $\hbar$, and $g \approx 2$. The so-called anomaly (the departure from the Dirac value) in QED is

$$\frac{(g-2)}{2} = 0.5\left(\frac{\alpha}{2\pi}\right) + \text{terms in } \alpha^2, \alpha^3, \ldots \tag{3.20}$$

where $\alpha = 1/137.06\ldots$ and the terms have been calculated up to those in α^4. The observed and predicted values are seen to agree precisely within the errors of less than one part in 10 million. It may be remarked that the errors shown for the theoretically predicted numbers are larger than those in the observed values, because they depend on the experimental uncertainty in the determination of α.

The success of gauge-invariant theories (QED in this case) can be compared with theories of the past not possessing gauge symmetry, which failed because they contained incurable divergence problems when calculations were made to high orders in the couplings involved. These divergent terms could in principle be cancelled, but only by introducing an indefinitely large number of arbitrary constants, thus losing any predictive power.

3.9 Superstrings

Unfortunately, no one has yet found a convincing way of extending the above ideas to gravity. A quantum theory of gravity does exhibit severe divergences, which can be greatly reduced but not totally eliminated in a supersymmetric version of the theory called supergravity (see Section 4.5 for a discussion of supersymmetry). In *superstring theory*, which embraces all the fundamental interactions, the pointlike particles (and their supersymmetric partners) responsible for these divergences are replaced by short (10^{-33} cm) *strings*, that is, of the order of the Planck length (1.12). The different elementary particles are described by the different modes of oscillation of such strings. Gravity appears then to be renormalizable, but only in 10-dimensional space–time (for fermions, or 26-dimensional space–time for bosons). The normal four dimensions of space and time we actually observe will result if all the extra dimensions are 'curled up' into the tiny extent of the string length, according to the original ideas of Kaluza and Klein in the 1920s. The weakness of the gravitational force is then ascribed to the fact that at normal energies, these extra dimensions are ineffective, and it is only when one reaches energies of the order of the Planck energy (or lengths comparable with the Planck length) that gravity becomes strong. Superstring theory does correctly predict the spin 2 of the graviton. Unfortunately, it also predicts a massive superpartner, the gravitino of spin 3/2. If this were produced, like the other elementary particles, in the hot early universe, its decay products would completely alter the predictions on nucleosynthesis of the light elements and ruin the close agreement with experiment, as described in Chapter 6. In summary, incorporating gravity with the other interactions is still an unsolved problem.

3.10 Gauge invariance in the electroweak theory

In QED, we saw that gauge invariance is associated with an infinite set of phase transformations of the wavefunction of the form

$$\psi \to \psi \exp\left[ie\alpha\left(x\right)\right] \tag{3.21}$$

These transformations are actually elements of a group called U(1), the 'U' standing for unitary, implying that the norm of the wavefunction is preserved in the transformation, and the '1' that it refers to a rotation in one dimension. In the electroweak theory, more complicated transformations, belonging to the SU(2) group are also involved. They are of the form,

$$\psi \to \psi \exp\left[ig\tau \cdot \Lambda\right] \tag{3.22}$$

Here the transformation involves the Pauli 2×2 matrices $\tau = (\tau_1, \tau_2, \tau_3)$ and describes rotations about the arbitrary vector Λ. The Pauli matrices were originally invented to describe spin 1/2 particles, the '2' in the nomenclature SU(2) referring to the dimension of the matrices, the 'U' indicating that the transformation is again unitary. The 'S' stands for 'special', SU(2) being a subgroup of U(2) in which the matrices are traceless. A fundamental difference between the transformations (3.21) and (3.22) is that U(1) is an Abelian group since $\alpha(x)$ is a scalar quantity. Thus the effect of two rotations in succession is independent of the order and $\alpha_1\alpha_2 - \alpha_2\alpha_1 = 0$, that is, the two operations *commute*. On the other hand, the group SU(2) is non-Abelian, involving the *non-commuting* Pauli operators, for example, $\tau_1\tau_2 - \tau_2\tau_1 = i\tau_3$.

The electroweak model was introduced by Glashow (1961), Weinberg (1967), and Salam (1967). It postulates four massless vector bosons; a triplet w^+, w^- and w^0 belonging to the SU(2) group and b^0 belonging to the U(1) group, that is, a system with SU(2) × U(1) symmetry. The neutral component w^0 mixes with the b^0, to form the photon γ and a neutral boson z^0, involving an arbitrary mixing angle θ_W. Finally, scalar bosons called *Higgs scalars* (after their inventor, Higgs (1964)) are postulated, to generate mass by self-interaction, as described below. Three of the four Higgs components are absorbed by the states w^+, w^-, and z^0, to form the massive vector bosons W^+, W^-, and Z^0 introduced in Chapter 1, while the photon γ remains massless.

Furthermore, although massive bosons are involved, the theory does remain renormalizable. The *weak and electromagnetic interactions are unified*, and the coupling of the W to leptons, specified by the coupling constant g in (3.22), is given by the relation $e = g \sin\theta_W$. (There are several numerical factors entering in the definition of g, which have arisen historically. The quantity g_w, which we introduced in (1.9) as $g_w^2 = G_F M_W^2$, is related to g by $g_w^2 = \sqrt{2}g^2/8$). The two unknown parameters in the model are the photon mass (zero) which has to be put in 'by hand', and the above mixing angle, which has been measured as $\sin^2\theta_W = 0.231 \pm 0.001$. The boson masses are then predicted in terms of the Fermi weak interaction constant G_F, e, and the mixing angle:

$$M_\mathrm{W} = \left[g^2 \frac{\sqrt{2}}{(8G_\mathrm{F})}\right]^{1/2} = \left[e^2 \frac{\sqrt{2}}{\left(8G_\mathrm{F}\sin^2\theta_\mathrm{W}\right)}\right]^{1/2} = \frac{37.4}{\sin\theta_\mathrm{W}}\,\mathrm{GeV}$$

$$M_Z = \frac{M_\mathrm{W}}{\cos\theta_\mathrm{W}} \tag{3.23}$$

The electroweak theory was vindicated by the discovery in 1973 of neutral weak currents, that is, the existence of Z^0 exchange as in Fig. 1.3, and by the observation of the W and Z bosons in 1983 (see Figs. 1.6 and 1.13). Note that, because the W and Z bosons are massive, compared with the zero mass of the photon, the SU(2) × U(1) symmetry of the model is broken by the Higgs mechanism of mass generation, but because the theory remains renormalizable, cross-sections and decay rates mediated by the bosons W and Z can be calculated exactly. All that is missing at the present time is the fourth Higgs component, which should exist as a physical particle. A lower limit on the mass is $M_\mathrm{H} > 100\,\mathrm{GeV}$. Finding the elusive Higgs is one of the prime objectives of experimental high-energy physics at the present time.

3.11 The Higgs mechanism of spontaneous symmetry breaking

We now discuss briefly the Higgs mechanism for spontaneous symmetry breaking in the electroweak theory. It is relevant to introduce it here, not only because it is an intrinsic part of the very successful electroweak theory, but also because a somewhat similar mechanism has been postulated in connection with the inflationary model of the early universe, which is discussed in Chapter 8.

As stated in Section 3.1, the equation for the Lagrangian energy density L of a field Φ in a quantum-mechanical system is written as

$$\frac{\partial}{\partial x_\mu}\left(\frac{\partial L}{\partial \Phi'}\right) - \frac{\partial L}{\partial \Phi} = 0 \tag{3.24}$$

where $\Phi' = \partial\Phi/\partial x_\mu$, Φ is the amplitude of the field particles and x_μ (with $\mu = 0, 1, 2, 3$) is the space–time coordinate (so in units $\hbar = c = 1, x_0 = t, x_1 = x, x_2 = y$, and $x_3 = z$). For free scalar particles of mass μ the Lagrangian function has the form

$$L = T - V = \left(\frac{1}{2}\right)\left(\frac{\partial\Phi}{\partial x_\mu}\right)^2 - \frac{\mu^2\Phi^2}{2} \tag{3.25}$$

which gives for the equation of motion in (3.24) the expression (known as the Klein–Gordon equation—see Appendix B):

$$\left(\frac{\partial^2}{\partial\mathbf{r}^2} - \frac{\partial^2}{\partial t^2} - \mu^2\right)\Phi = 0$$

With the substitution of the operators $E = -i\partial/\partial t, \mathbf{p} = -i\partial/\partial\mathbf{r}$, this becomes the usual relativistic relation between total energy, three-momentum, and mass:

$$-|\mathbf{p}|^2 + E^2 - \mu^2 = 0$$

Suppose now that we are dealing with scalar particles *which interact with each other*. This means adding an extra term to (3.25) which is of the form Φ^4 (odd powers are excluded because of symmetry required in the transformation $\Phi \to -\Phi$, and powers higher than the fourth by the requirement of renormalizability). So the modified Lagrangian is written as

$$L = \left(\frac{1}{2}\right)\left(\frac{\partial\Phi}{\partial x_\mu}\right)^2 - \left(\frac{1}{2}\right)\mu^2\Phi^2 - \left(\frac{1}{4}\right)\lambda\Phi^4 \tag{3.26}$$

where λ is a dimensionless constant representing the coupling of the four-boson vertex. The minimum of the potential V occurs when $\partial V/\partial\Phi = 0$, that is, when

$$\Phi\left(\mu^2 + \lambda\Phi^2\right) = 0 \tag{3.27}$$

If $\mu^2 > 0$, the situation for a massive scalar field particle, then $\Phi = \Phi$ (min) when $\Phi = 0$, as is the usual case with the vacuum state having $V = 0$. However, it is also possible to consider the case $\mu^2 < 0$, where $\Phi = \Phi$ (min) when

$$\Phi = \pm v = \pm\left(\frac{-\mu^2}{\lambda}\right)^{1/2} \tag{3.28}$$

In this case, the lowest energy state has Φ finite, with $V = -\mu^4/4\lambda$ and instead of being zero, V is everywhere a non-zero constant. The quantity v is called the vacuum expectation value of the field Φ. The situation is illustrated in Fig. 3.5. The minimum at $\Phi = 0$ is referred to as the *false vacuum* and that at $\Phi = \pm v$ as the *true vacuum*, being the lowest energy state.

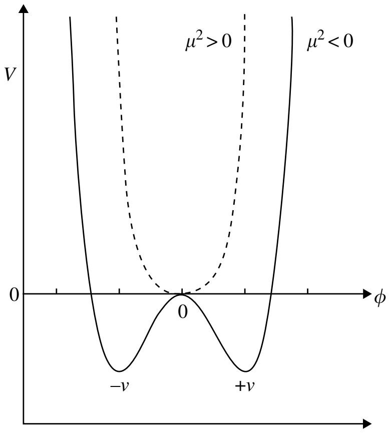


Fig. 3.5 The potential in (3.26) as a function of Φ, the value of a one-dimensional scalar field, for the cases $\mu^2 > 0$, and $\mu^2 < 0$.

In the context of electroweak interactions, one is concerned with *small* perturbations about the energy minimum, so the field variable Φ should be expanded, not about zero but about the chosen vacuum minimum ($+v$ or $-v$

in the above example). If one writes

$$\Phi = v + \sigma(x) \tag{3.29}$$

where σ is the value of the extra field over and above the constant and uniform value v, then substituting into (3.25) one gets

$$L = \left(\frac{1}{2}\right)\left(\frac{\partial\sigma}{\partial x_\mu}\right)^2 - \lambda v^2\sigma^2 - \left(\lambda v\sigma^3 + \frac{\lambda\sigma^4}{4}\right) + \text{constant} \tag{3.30}$$

where the constant terms involve powers of v only. The third term represents the interaction of the σ field with itself. The second term, when compared with the potential in (3.26), is clearly a mass term, with a value for the mass of

$$m = \sqrt{2\lambda v^2} = \sqrt{-2\mu^2} \tag{3.31}$$

So, by making a perturbation expansion about either of the minima $+v$ or $-v$, a *positive real mass* has appeared. Note that the expansion has to be made about *one* of the two minima. Of course, once this is done, the symmetry of Fig. 3.5 is broken. Such a behaviour is called *spontaneous symmetry breaking*. Many examples exist in physics. A bar magnet heated above the Curie point has its elementary magnetic domains pointed in random directions, with zero net moment, and the Lagrangian is invariant under rotations of the magnet in space. On cooling, the domains will set in one particular direction, that of the resultant moment, and the rotational symmetry is spontaneously broken.

The treatment above was of a one-component scalar field. For the more general case of a complex scalar field, $\Phi_1 + i\Phi_2$, the two points $\pm v$ in Fig. 3.5 are replaced by all the points on a circle of radius v obtained by rotating the diagram about a vertical axis. However, the principle of obtaining a real mass associated with the lowest energy 'true' vacuum state by spontaneously breaking the symmetry of the potential remains as before.

The next step is to replace the derivative $\partial/\partial x_\mu$ in (3.30) by the covariant derivative analogous to that in (3.19) but extended to include both the U(1) and SU(2) transformations in (3.21) and (3.22). When this is done, one obtains relations for the squares of the masses of the W and Z bosons as in (3.23), and also in terms of the Higgs vacuum term v. The measured values of the boson masses give $v = 246$ GeV, which is thus the scale of the electroweak symmetry breaking. However, the Higgs mass is not directly predicted by the theory, but it should have a mass of the order of the electroweak scale and in any case less than 1 TeV.

3.12 Running couplings: comparison of electroweak theory and quantum chromodynamics with experiment

In Section 3.8 it was noted that in gauge theories, perturbation calculations giving finite answers can be carried out to any order in the coupling constant. However, there is a practical limit. For example, in calculating the $(g - 2)$ correction to the electron magnetic moment, there are already 72 Feynman diagrams to be summed over for the term in α^3. The situation is only saved by the smallness of $\alpha \sim 1/137$ and the uncertainty in its experimental value, which together make higher-order terms, in α^5 or higher, unimportant or irrelevant in comparing the predicted $(g-2)$ with experiment. Since for the strong interquark interactions, the coupling α_s is much greater than α, the complications in quantum chromodynamics (QCD) calculations would be much worse.

Fortunately, to a good level of approximation (called the leading log approximation) it is possible to replace the perturbation series by a single term, an *effective coupling* which is not constant but depends on the four-momentum transfer q in the process considered. For the electromagnetic interaction, the formula is

$$\alpha\left(q^2\right) = \frac{\alpha\left(\mu^2\right)}{\left[1-(1/\pi)\,\alpha\left(\mu^2\right)\ln\left(q^2/\mu^2\right)\right]} \tag{3.32}$$

The formula relates the coupling at one momentum transfer q to that at another momentum μ (incidentally avoiding any problem of the coupling at infinite momentum). The effective coupling is *increasing* with the energy scale. Why is that? Consider a test charge immersed in a dielectric (see Fig. 3.6). The atoms of the dielectric become polarized, and this produces a *shielding* effect, so that the potential due to the test charge at distances large compared with atomic dimensions is less than it would be without the dielectric. So the effective value of the test charge is reduced at large distances but increases as one probes in to smaller distances or equivalently to larger momentum transfers. A similar effect is possible even in a vacuum, since the test charge is continually emitting and reabsorbing virtual pairs—the process called *vacuum polarization* described before—and equation (3.32) gives the quantitative evaluation of this shielding effect or running of the coupling.

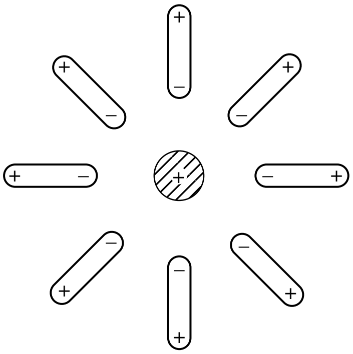


Fig. 3.6

Example 3.3 *The electromagnetic coupling parameter* $\alpha \approx 1/137$ *at low momentum transfers,* $\mu \sim 1$ *MeV. Calculate the value of* α *at the electroweak scale (*$q \sim 100$ *GeV) and at the GUT scale (*$q \sim 3 \times 10^{14}$ *GeV).*

From equation (3.32) we have

$$\frac{1}{\alpha\left(q^2\right)} = \frac{1}{\alpha\left(\mu^2\right)} - \frac{1}{\pi}\ln\left(\frac{q^2}{\mu^2}\right)$$

and substituting for the values of q^2, we find $1/\alpha = 137 - 7.3 \sim 129$ at the electroweak scale, and $1/\alpha = 137 - 25.6 \sim 111$ at the GUT scale. In the latter case, the change is so large that next to leading order terms (in α^2) probably need to be included in (3.32), which is the so-called leading log approximation, applying for small changes to the coupling.

For strong interactions (QCD) it turns out that, in addition to the shielding effect of fermion (quark) loops there is also an *anti-shielding* effect, because of the loops containing gluons and the (longitudinal component of) gluon–gluon coupling as shown in Fig. 3.7. This coupling increasingly ‘spreads’ the strong colour charge at the larger values of q^2. In this case, the dependence of the

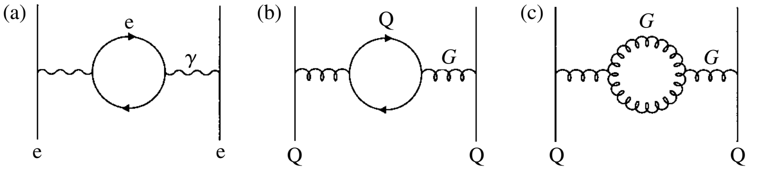


Fig. 3.7 Diagrams involving vacuum polarization effects (a) in QED where loops contain fermions only, and (b) and (c) in QCD, where loops contain both fermions and gluons, and the gluon–gluon coupling (for longitudinal gluon components) produces an anti-shielding effect.

strong coupling α_s is found to be

$$\alpha_s\left(q^2\right) = \frac{\alpha_s\left(\mu^2\right)}{\left[1 + B\alpha_s\left(\mu^2\right)\ln\left(q^2/\mu^2\right)\right]} = \frac{1}{\left[B\ln\left(q^2/\Lambda^2\right)\right]} \tag{3.33}$$

where $B = 7/4\pi$ and $\Lambda^2 = \mu^2 \exp\left[-1/B\alpha_s\left(\mu^2\right)\right]$, so that α_s *decreases* with increasing q^2. In the limit of very high q^2, this means that $\alpha_s \to 0$, a phenomenon known as *asymptotic freedom*, which developed through a long history which we now describe briefly.

3.12.1 From the parton model to QCD

The theory of strong interactions (QCD) had its origins in experiments on the deep inelastic scattering of high-energy leptons by nucleons. The pointlike leptons were employed to probe the structure of nucleons. In 1968, electron scattering experiments at Stanford found the first evidence for quarks as real dynamical objects. The inelastic cross-sections were found to be large and described by structure functions which were only weakly dependent on the four-momentum transfer, q^2 (Friedman and Kendall 1972). This was in contrast with the elastic electron–nucleon cross-sections, described by the pointlike Rutherford cross-section (1.23) multiplied by so-called form factors falling off rapidly with increasing q^2. So the weak q^2 dependence in the inelastic process was the signal of *elastic* scattering by *quasi-free, pointlike* constituents called *partons* (Feynman 1969), subsequently to be identified with quarks and gluons.

Figure 3.8 depicts an electron–nucleon collision in a reference frame where the target nucleon has very large four-momentum P (the 'infinite momentum frame'). In this frame, all particle masses can be neglected in comparison with their energies and momenta, and the constituent partons travel in a parallel beam, because transverse momenta are also negligible. Suppose now that the electron scatters *elastically* by transferring four-momentum q (via a photon) to one of the partons of mass m, carrying a fraction x of the nucleon four momentum. Then the four-momentum squared

$$\left(xP + q\right)^2 = m^2 \approx 0$$

If q is large, then $x^2P^2 = x^2M^2 \ll q^2$, where M is the proton mass, so $2xPq + q^2 = 0$ and

$$x = \frac{|q^2|}{2Pq} = \frac{|q^2|}{2M\nu}$$

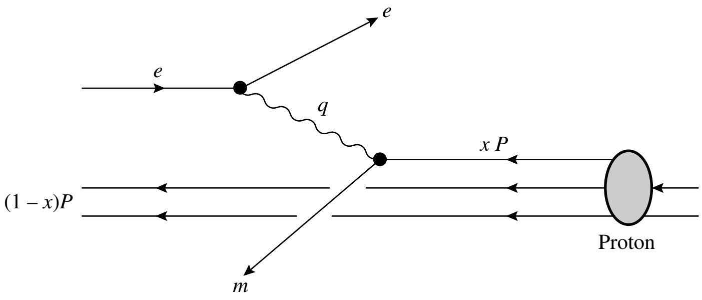


Fig. 3.8 Collision of a high energy electron with a parton carrying 4-momentum xP, viewed in the "infinite momentum frame" of the parent proton. q is the 4-momentum transfer.

where we recall that q^2 is negative in a scattering process. Here, the invariant product Pq has been evaluated in the laboratory frame, where the proton is at rest (with zero three momentum), and the energy components of P and q are just the mass M and the kinetic energy transfer ν. Again, ν is assumed to be large compared with any of the masses involved.

The experimental cross-section measures the so-called structure function $F(x)$, which is just the distribution in the momentum fraction x carried by the partons, and the q^2 dependence is all in the dimensionless combination $q^2/2M\nu$.

Obviously, we do not see partons emerging in the final state. The scattered and unscattered partons must recombine in a final state interaction, forming hadrons. This last is a slow process, compared with the timescale of the original electron–parton collision, so that the cross-section for the process depends first and foremost on the kinematics of the initial collision. We can just forget all the complications of how the partons rearrange themselves to make the messy, multi-hadron final state (see Fig. 3.10(a)).

What are these partons? The Stanford experiments with 25 GeV electrons and spectrometer detectors measured the scattered electron and the 'structure function' called $F_2(x)$. There are actually two functions, $F_2(x)$ for the electric and the other, $F_1(x)$ for the magnetic contributions—and it was shown from the ratio that the partons were fermions, spin $(1/2)\hbar$. A little later (1972) experiments at CERN in the large heavy liquid bubble chamber Gargamelle, also measured structure functions, using neutrino and antineutrino beams (there are three structure functions in this case, but with both neutrino and antineutrino events, one can eliminate one of them). The *shape* of the $F_2(x)$ function for nucleons (i.e. a proton/neutron average) measured with neutrinos was *the same* as that found with electrons. Despite the fact that the two experiments were totally different, in technique as well as in type of fundamental interaction, they were seeing the same elementary parton structure. Figure 3.9 shows the early results comparing the two cross-section measurements (Perkins 1972). After taking into account the difference in couplings in the two cases $\left(\alpha^2/q^4 \text{ and } G_{\mathrm{F}}^2 s/\pi\right)$ one observed the simple result that F_2(electron) was equal to $(5/18)\, F_2$(neutrino). The electron scattering will be weighted by the square of the parton charges, and with equal numbers of u and d quarks, that is clearly $\left[(2/3)^2 + (1/3)^2\right]/2$. So the observation of the $5/18$ ratio was a proof that the partons were the long-sought fractionally charged quarks.

The neutrino cross-sections also depended on whether the scattering was from a quark or antiquark, and thus showed that in addition to the 3 'valence'

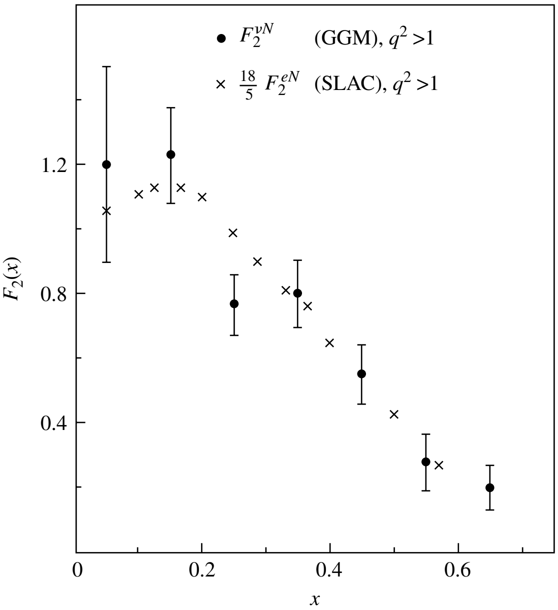


Fig. 3.9 Early data on the structure function $F_2^{\nu N}(x)$ measured from neutrino and antineutrino scattering by nucleons at CERN in the Gargamelle bubble chamber, compared with 18/5 times the value of $F_2^{eN}(x)$ measured in ep and ed scattering at SLAC. This was the first evidence (1972) for fractionally charged quarks as dynamical constituents inside the nucleon. Note that the integral over x, the fractional quark momentum, is about 0.5. The remaining momentum is ascribed to gluon constituents.

quarks, the nucleon in these collisions also contained some 15% of virtual quark–antiquark pairs, as we might have expected from diagrams of the same type as in Fig. 3.4(b). Integrating the curve in Fig. 3.9 one can only account for about 50% of the nucleon momentum; the rest is ascribed to gluon constituents.

The above discussion has treated the quark–partons as free particles, but this picture (applying literally at infinite momentum) cannot be correct, and deviations from it at finite collision energies were predicted in perturbative QCD in 1973 (Politzer 1973). These deviations were first observed and quantified in bubble chamber neutrino experiments at CERN in 1978 (Bosetti *et al.* 1978, 1982), which measured a first value for the parameter $\Lambda \sim 200$ MeV in the expression (3.33) for the q^2 dependence of α_s. Figure 3.10 shows a typical neutrino event, and Fig. 3.11 the recent values for the q^2 dependence of the strong coupling.

3.12.2 Testing the Standard Model

The running of the couplings is important in performing precision fits of data on electroweak interactions to the Standard Model. The data come from measurements at giant e^+e^- colliders of the W and Z boson masses and widths, the forward–backward asymmetry in the decays of these bosons to leptons and to hadrons (*via* quark pairs), and the cross-sections for neutrino and antineutrino scattering on electrons and on nucleons. The different quantities or processes, when evaluated theoretically, will contain different contributions from radiative corrections. Figure 3.12 shows examples of how such corrections can affect α (as described earlier in this section) or the W boson mass.

Figure 3.13 shows how the Standard Model is tested. Some quantities, such as $M_Z = 91.189 \pm 0.001$ GeV, have been measured very accurately. Theoretically, the radiative corrections to the Z boson mass, and to the quantity $\sin^2\theta_w$,

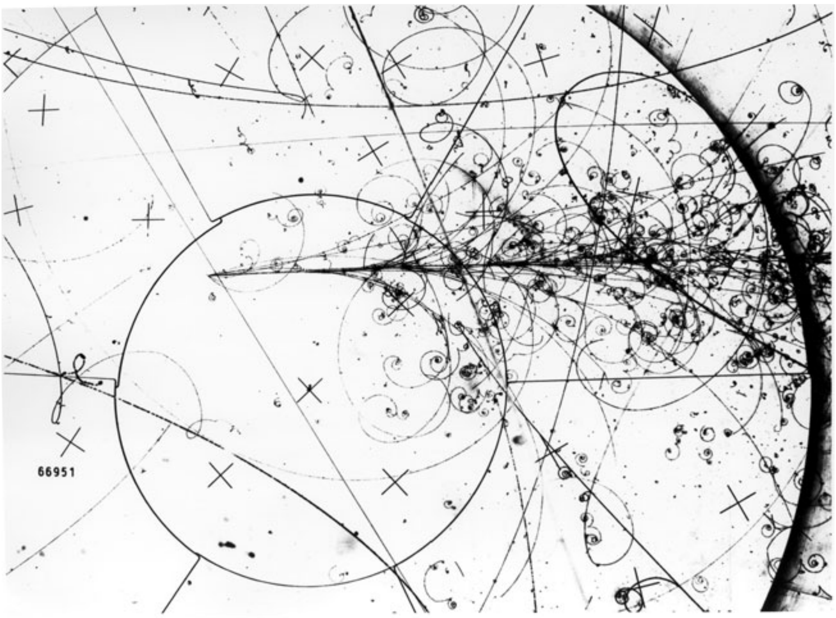


Fig. 3.10 Example of deep inelastic neutrino–nucleon collision in the BEBC bubble chamber (CERN) filled with a liquid hydrogen–neon mixture. When the chamber is expanded, bubbles form along the tracks of charged particles. In this picture, a muon–neutrino of 200 GeV energy enters horizontally from the left, and transforms to a muon in an elastic collision with a quark: $\nu_\mu + d \to u + \mu^-$, with a value of $q^2 \approx 75\,\text{GeV}^2$. The muon forms the rather straight track at 2.30 pm. The other tracks are due to hadrons (pions) produced when the quarks interact in the final state. Neutral pions decay to γ-rays which generate electron–positron pairs and cascades in the heavy liquid (see Section 9.6). Particle momenta are measured from track curvature in the applied (5 kG) field. Analysis of several hundred such events gave the first quantitative evidence in support of perturbative QCD.

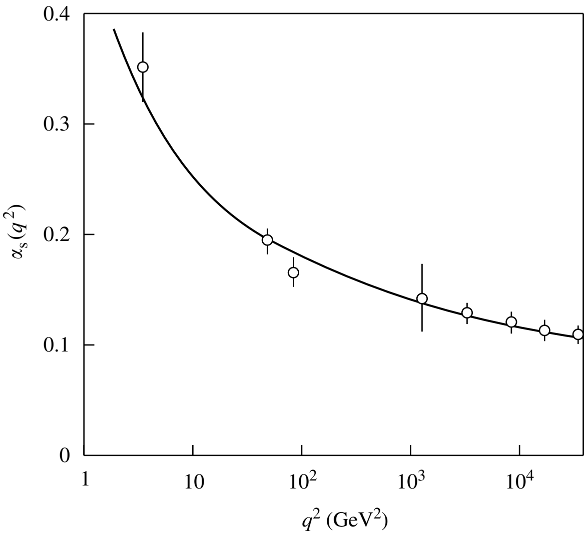


Fig. 3.11 Variation of the QCD 'running coupling' with q^2, the data coming from a variety of sources, including the τ lepton width, inelastic lepton–nucleon scattering, upsilon ($= b\bar{b}$) decays, Z^0 width, and event shapes and widths in the process $e^+e^- \to$ hadrons. The curve is the prediction for $\Lambda = 200$ MeV in (3.33).

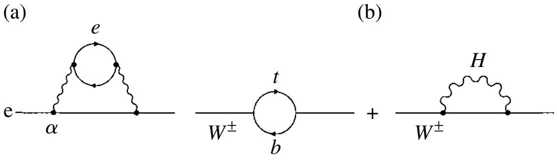


Fig. 3.12 Loop diagrams indicating radiative corrections (a) to α from a virtual fermion loop and (b) to the mass of the W or Z bosons from loops containing a virtual top quark or a Higgs scalar.

both depend, for example, on the mass of the top quark. The figure shows the expected variation in $\sin^2\theta_w$ with M_{top} for the observed value of M_Z. One can also determine $\sin^2\theta_w$ in other processes with different radiative corrections. And the mass of the top quark has also been determined by direct experiment, rather than from a radiative correction. The question then is whether all the data

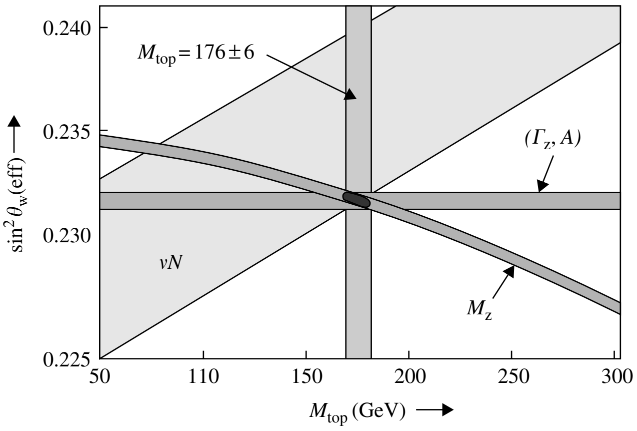


Fig. 3.13 Values of the electroweak mixing angle versus top quark mass, computed from the radiative corrections (as in Fig. 3.12) to various quantities, for example, the Z boson mass, neutrino–nucleon scattering cross-sections (for neutral versus charged currents), and the widths and asymmetries in Z decays. The best fit, where the various curves intersect, is in excellent agreement with the top quark mass measured directly.

put together can give a unique fit to the model, with a set of best-fit parameters? Clearly this is so; the best fit is indicated by the dark area at the centre of the plot. Although in this plot, the Higgs mass was assumed to be 300 GeV, this quantity can also be determined in the fit, although not very precisely because the radiative corrections depend only logarithmically on the Higgs mass. When this is done, a rather light Higgs mass, $M_H < 160$ GeV is indicated.

3.13 Vacuum structure in gauge theories

Non-Abelian gauge theories such as the electroweak theory described in the foregoing can possess complex vacuum structures, concerned with non-perturbative processes. These vacuum states are characterized by different topologies, corresponding to different additive quantum numbers (lepton and baryon number), and are separated by potential barriers. Transitions between the vacua can be made by quantum-mechanical tunnelling in so-called *instanton* processes, resulting in changes of baryon or lepton number ('t Hooft 1976). An analogy can be made with the alpha particle decay of a radioactive nucleus through the Coulomb barrier, resulting in a change $\Delta B = 4$ in baryon number. In the low temperature approximation, such processes are completely suppressed by the negligibly small value of the barrier penetration probability between adjacent vacua. However, at high temperatures, such as could occur at a very early stage in the expansion of the universe, the thermal energy may be enough for jumps over, rather than through, the barrier, in so-called *sphaleron* processes, which involve a 12-lepton vertex (three quarks and one lepton, for each of the three generations). As discussed in more detail in Chapter 6, these processes have been proposed as one possible mechanism contributing to the baryon asymmetry of the universe.

3.14 CPT theorem and CP and T symmetry

As will be discussed in Chapter 6, the development of the observed baryon asymmetry of the universe, which is assumed to have started off in a state

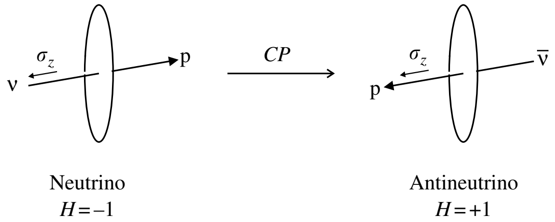


Fig. 3.14 The operation P on an LH neutrino transforms into an RH neutrino state, which is not observed. The C-operation on an LH neutrino state transforms into an LH antineutrino state, which is also not observed. The combined CP operation however transforms an LH neutrino into an RH antineutrino, which *is* observed.

of matter–antimatter symmetry, can only be understood in terms of out-of-equilibrium processes and violation of CP symmetry, which is described in this section.

The operations of charge conjugation C, of spatial inversion P and of time reversal T are connected through the very important *CPT theorem.* This states that *all* interactions are invariant under the three operations C, P, and T taken in any order. The theorem predicts that the masses, magnetic moments, lifetimes, etc. of particles and antiparticles should be identical, a prediction which is verified to very high accuracy. For example, the difference in masses of the neutral kaon K^0 and its antiparticle $\bar{K}^0$ is less than 1 in 10^{19}, while the difference in absolute values of the magnetic moments of the positron and electron is less than 1 part in 10^{12}. The CPT theorem also predicts the spin-statistics relation, that integral and half-integral spin particles obey Bose–Einstein and Fermi–Dirac statistics respectively.

While CPT invariance is, as far as we know, universal, CP and T symmetries are not. Let us recall from (3.9) and Fig. 3.3 that while the weak interactions are not invariant under C or under P, the operation CP does transform a LH neutrino state into the RH state of its charge conjugate, the antineutrino—see Fig. 3.14. In fact, for a time it was thought that the CP symmetry might be universal, but then the evidence for CP violation was observed in the decay of neutral kaons, as we now discuss.

3.15 CP violation in neutral kaon decay

The kaons are the lightest mesons formed from the combination of a strange quark or antiquark with a non-strange antiquark or quark. They are produced in strong interactions of hadrons and occur in four states, all of spin-parity $J^{\mathrm{P}} = 0^-$ and with masses of 0.494 GeV/c^2 for $K^+(= u\bar{s})$ and $K^-(= \bar{u}s)$, and 0.498 GeV/c^2 for $K^0(= d\bar{s})$ and $\bar{K}^0(= \bar{d}s)$. The states with a strange quark have $S = -1$, while those with a strange antiquark have $S = +1$. All the kaon states are unstable. The charged kaons, being particle and antiparticle, have the same mean lifetime of 12.4 ns. For the neutral kaons, however, two different lifetimes are observed. The state called K_S has $\tau = 0.089$ ns and that called K_L has $\tau = 51.7$ ns (the subscripts standing for ‘short’ and ‘long’). The existence of two lifetimes arises because the decaying states the experimentalist detects are superpositions of K^0 and $\bar{K}^0$ amplitudes. This mixing occurs through virtual 2π and 3π intermediate states and involves a *second-order weak interaction of*

$\Delta S = 2$:

$$\begin{tikzcd}
 & 2\pi \arrow[dr] & \\
K^0 \arrow[ur] \arrow[dr] & & \bar{K}^0 \arrow[dl] \\
 & 3\pi &
\end{tikzcd}$$

First, we can form CP eigenstates from the neutral kaon states as follows:

$$K_S = \sqrt{\frac{1}{2}}\left(K^0 + \bar{K}^0\right) \quad \mathrm{CP} = +1$$

$$K_L = \sqrt{\frac{1}{2}}\left(K^0 - \bar{K}^0\right) \quad \mathrm{CP} = -1 \tag{3.34}$$

where, since the kaons have spin zero, the operation CP on the wavefunction has the same effect as that of charge conjugation, C. On taking into account the negative intrinsic parity of the pion mentioned in Section 3.4, the decay modes will be $K_S \rightarrow 2\pi$ where the final state consists of two pions in an S-state with CP $= +1$ and $K_L \rightarrow 3\pi$ with CP $= -1$. Thus, while the neutral kaons are *produced* as eigenstates of strangeness, K_0 and $\bar{K}^0$, they *decay* as superpositions of these states which are actually eigenstates of CP.

In 1964, it was found by Christenson *et al.* that the above states were in fact *not* pure CP eigenstates. If we denote a pure CP $= +1$ state by K_1, and a pure CP $= -1$ state by K_2, the K_L and K_S amplitudes are written as

$$K_S = N\,(K_1 - \varepsilon K_2)$$

$$K_L = N\,(K_1 + \varepsilon K_2) \tag{3.35}$$

where the normalizing factor $N = (1 + |\varepsilon|^2)^{-1/2}$ and $\varepsilon \approx 2.3 \times 10^{-3}$ is a small parameter quantifying the level of CP violation. The experiment commenced with a beam of K^0 generated in a strong interaction. After coasting for several K_S mean lives, the experimenters were left with a pure K_L beam. It was observed that a small proportion of the K_L decays were to a two-pion state, with $CP = +1$ (see Fig. 3.15).

CP violation is also demonstrated in the leptonic decay modes of K_L. If we denote the rate for $K_L \rightarrow e^+ + \nu_e + \pi^-$ by R^+, and for $K_L \rightarrow e^- + \nu_e + \pi^+$ by R^-, then it is observed that

$$\Delta = \frac{\left(R^+ - R^-\right)}{\left(R^+ + R^-\right)} = (3.3 \pm 0.1) \times 10^{-3} \tag{3.36}$$

One of the most striking features of the universe is the very large asymmetry between matter and antimatter, as will be discussed in Sections 6.4 and 6.5. There we note that CP-violating interactions are necessary in order to generate a baryon–antibaryon asymmetry. The result (3.36) emphasizes that CP violation is actually required to differentiate unambiguously between matter and antimatter on a cosmic scale. Here on Earth we define the positron of antimatter as having a positive charge and the electron as negative. But these are just names and

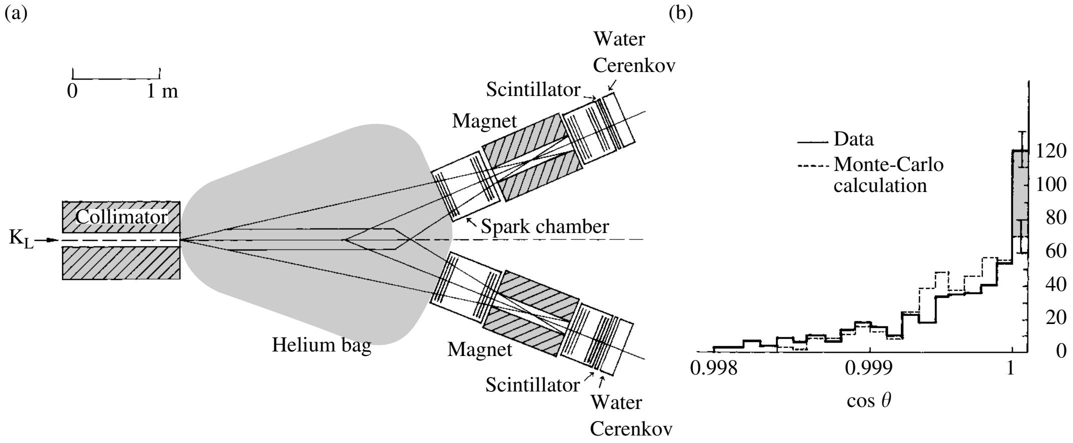


Fig. 3.15 Arrangement of the experiment by Christenson *et al.* (1964), demonstrating the CP violating $K_L \to \pi^+\pi^-$ decay. The charged products of the decays were analysed by two magnet spectrometers instrumented with spark chambers and scintillators. The rare two-pion decays are distinguished from the common three-pion decays by requiring that the two-pion invariant mass should be consistent with the kaon mass, and that the resultant vector momentum of the two pions should be in the beam direction. The distribution in $\cos\theta$ is that expected for three-pion decay, plus some 50 events collinear with the beam and attributed to the rare two-pion mode.

what we define as positive or negative charge is quite arbitrary. All physical results would have been the same if we had defined the electron as positive and the positron as negative. So we need an unambiguous way of defining what we call matter and antimatter to an intelligent being in a far corner of the universe. CP violation in neutral kaon decay now provides the answer. The positron is defined as that charged lepton which is more prolific (by 0.3%) in the long-lived K_L decay.

Example 3.4 *If the annihilation of proton and antiproton proceeds through an S-state, show that $p\bar{p} \to K_1 + K_2$ can occur, but not $p\bar{p} \to K_1K_1$ or K_2K_2, where K_1 and K_2 are eigenstates of $CP = +1$ and -1 respectively.*

A proton–antiproton system with total angular momentum L and total spin S has symmetry $(-1)^{L+S}$ under interchange of space and spin coordinates. But this is equivalent to charge conjugation or particle–antiparticle conjugation, leaving space and spin alone. Hence the system has $C = (-1)^{L+S}$ and parity $P = (-1)^{L+1}$, taking account of the opposite parities of particle and antiparticle. Hence the initial state of proton and antiproton has

$$\mathrm{CP} = (-1)^{2L+S+1} = (-1)^{S+1} \quad \text{for all } L \text{ values}$$

Let J be the total angular momentum of the two kaons, where $|L + S| \geq J \geq |L - S|$. Measured in their rest-frames, the K_1 has $\mathrm{CP} = +1$ and the K_2 has $\mathrm{CP} = -1$. If the orbital angular momentum of the pair is J, this introduces a factor $(-1)^J$ for the parity. Hence in the final state,

$$\text{For } 2K_1 \quad \mathrm{CP} = (+1)\,(+1)\,(-1)^J = (-1)^J$$

$$\text{For } 2K_2 \quad \mathrm{CP} = (-1)\,(-1)\,(-1)^J = (-1)^J$$

$$\text{For } K_1 + K_2 \quad \mathrm{CP} = (-1)\,(+1)\,(-1)^J = (-1)^{J+1}$$

For annihilation from an *S-state*, $L = 0$ and $J = S$, so the initial state has $\mathrm{CP} = (-1)^{J+1}$ where $J = 0$ or 1. Thus annihilation to $K_1 + K_2$ is allowed and $2K_1$ or $2K_2$ is forbidden.

For annihilation from a *P-state*, $L = 1$ and if $S = 1, J = 0$, 1, or 2. In this case, $\mathrm{CP} = +1$ in the initial state so that $J = 0$ or 2 allows $2K_1$ or $2K_2$ in the final state, while if $J = 1$ only $K_1 + K_2$ is allowed. If $S = 0, J = 1$, the initial value of $\mathrm{CP} = -1$ and only the states $2K_1$ or $2K_2$ are allowed.

Experimentally, it is observed that for annihilation at rest only K_1K_2 is observed, as expected if an $L = 0$ state is involved.

3.16 CP violation in the Standard Model: the CKM matrix

There are in fact *two* sources of CP violation in neutral kaon decay. First, the states (3.35) with definite lifetimes are not pure CP eigenstates. This is known as *indirect* CP violation, occurring in the mass eigenstates themselves through a second-order transition of $\Delta S = 2$. But also, CP violation occurs in the actual decay process, which of course involves a first-order $\Delta S = 1$ transition. This is known as *direct* CP violation. It happens that in neutral kaon decay, the direct CP violating amplitude ε' is very small compared with the indirect amplitude ε. Indeed, it took more than 30 years from the first observation of CP violation to establish the existence of the direct process and measure it reliably. The ratio $\varepsilon'/\varepsilon = (16.6 \pm 1.6) \times 10^{-4}$. The Standard Model of particle physics makes some predictions about the level of direct CP violation. To introduce this, let us go back to the Fermi coupling in the weak interactions. The *leptons* are coupled to the $W^{\pm}$ mediating boson via a universal coupling specified by the Fermi constant G_F—see (1.9). However, for the *quarks*, the coupling to the $W^{\pm}$ is to weak interaction eigenstates which are *admixtures of flavour eigenstates*. The quark doublets analogous to the lepton doublets

$$\begin{pmatrix} \nu_e \\ \mathrm{e}^- \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \nu_\mu \\ \mu^- \end{pmatrix}$$

are written

$$\begin{pmatrix} u \\ d' \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} c \\ s' \end{pmatrix}$$

where

$$\begin{aligned} d' &= d\cos\theta_c + s\,\sin\theta_c \\ s' &= -d\sin\theta_c + s\,\cos\theta_c \end{aligned} \qquad \text{or} \qquad \begin{pmatrix} d' \\ s' \end{pmatrix} = \begin{pmatrix} \cos\theta_c & \sin\theta_c \\ -\sin\theta_c & \cos\theta_c \end{pmatrix} \begin{pmatrix} d \\ s \end{pmatrix} \tag{3.37}$$

The mixing angle $\theta_c = 12.7°$ is called the Cabibbo angle. Thus for neutron decay, which in quark language is written as $d \rightarrow u + \mathrm{e}^- + \bar{\nu}_e$, the coupling is $G_F \cos\theta_c = 0.975\ G_F$, while for the decay of a strange Λ-hyperon, $\Lambda \rightarrow \mathrm{p} + \mathrm{e}^- + \bar{\nu}_e$, or in quark symbols $s \rightarrow u + \mathrm{e}^- + \bar{\nu}_e$, the coupling is $G_F \sin\theta_c = 0.22\ G_F$. Here, we have purposely included only two of the three lepton and quark doublets in the above equation. When one includes all three families, that is, the doublets (ν_τ, τ^-) and (t, b), the transformation replacing (3.37) will be a 3 × 3 matrix called the CKM matrix after its proponents Cabibbo(1963) and Kobayashi and Maskawa (1972).

This is written as

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = V_{\text{CKM}} \begin{pmatrix} d \\ s \\ b \end{pmatrix}$$

where the absolute magnitudes of the (generally complex) elements of the matrix are

$$V_{\text{CKM}} = \begin{vmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{vmatrix} \approx \begin{vmatrix} 0.975 & 0.221 & 0.004 \\ 0.221 & 0.975 & 0.039 \\ 0.008 & 0.038 & 0.999 \end{vmatrix} \tag{3.38}$$

The extreme off-diagonal elements V_{td} and V_{ub} are very small and not well determined.

The important point is that an $N \times N$ matrix obviously contains $N\ (N-1)/2$ pairs, that is, Euler (mixing) angles, equal to 3 for $N = 3$ (compared with one for $N = 2$); and $(N-1)\ (N-2)/2$ arbitrary, non-trivial phases (i.e. one for a 3 × 3 matrix and none for a 2 × 2). This phase δ in the CKM matrix enters the wavefunction as $\exp\ [i(\omega t + \delta)]$, which is not invariant under time reversal $t \to -t$. So this is a possible T violating, or equivalently CP violating, amplitude in the Standard Model. The existence of this phase implies that some of the elements of the CKM matrix must be complex. Since on the hypothesis of universal Fermi coupling, the matrix V is unitary, the off-diagonal elements of the product $V * V$ must be zero. So, for example, multiplying the top row of the complex transpose matrix $V*$ by the last column of V, we get for the right-hand top corner element of the product matrix

$$V_{ud}^* V_{ub} + V_{cd}^* V_{cb} + V_{td}^* V_{tb} = 0 \tag{3.39}$$

and this can be plotted as a 'unitarity triangle' in the complex plane as shown in Fig. 3.16. The three angles α, β, and γ can be found by measurements on the decays of neutral B-mesons, that is the quark–antiquark combinations $b\bar{d}$ and $d\bar{b}$, known as B_d^0 and $\bar{B}_d^0$, and the combinations $b\bar{s}$ and $\bar{b}s$, called B_s^0 and $\bar{B}_s^0$. For B-meson decays, the direct CP violation process is dominant over the indirect. Pair production of $B-$mesons in enormous quantity has been achieved at electron–positron colliders especially built for the purpose and called 'B-factories' at Stanford, USA and KEK, Japan. Unlike the neutral kaons, the more massive B-mesons have very many decay modes. The level of CP violation is obtained by measuring the difference in decay rates of B^0 and $\bar{B}^0$ to the small fraction of the decay modes that are common to both. For example, the decays B_d^0 to ψK_s, where the ψ is the ground state $c\bar{c}$ meson resonance, measure sin 2β, while decays to $\pi^+\pi^-$ measure sin 2α, and B_s^0 to ρK_s measures sin 2γ. The object of these measurements is to determine whether the observed rates of CP violation are consistent with the constraints of the Standard Model. Although none has so far been observed, any departure from the unitarity relation (3.39)—non-closure of the triangle in Fig. 3.16—would be a signal of new physics beyond the Standard Model.

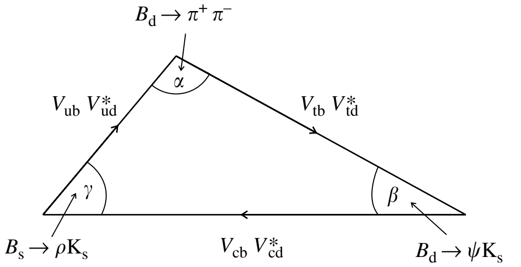

$B_d \rightarrow \pi^+ \pi^-$
α
$V_{ub} V^*_{ud}$
$V_{tb} V^*_{td}$
γ
β
$B_s \rightarrow \rho K_s$
$V_{cb} V^*_{cd}$
$B_d \rightarrow \psi K_s$

Fig. 3.16 'Unitarity triangle' showing the angles α, β, and γ determined by measurement of CP violation in neutral B-meson decays. Any departure from closure of this triangle, that is violation of the unitarity relation (3.39), would indicate new physics beyond that of the Standard Model. Current values are $\alpha = 99 \pm 11^0; \beta = 20 \pm 1^0; \gamma = 63 \pm 13^0$, with a sum of 182 ± 18^0 (Yao *et al.* 2006).

The important question for astrophysics is whether the level of CP violation included in the Standard Model is sufficient to account for the observed CP violation and matter–antimatter asymmetry on a cosmological scale. Current thinking is that the matter asymmetry must have been generated in the very early universe, at energies far in excess of those associated with the Standard Model and laboratory experiments. CP violation in K^0 and $B-$meson decays would seem to be irrelevant to CP violation on the scale of the early universe.

3.17 Summary

- Symmetries and invariance principles give rise to conservation rules. Invariance of the Lagrangian function under a global phase transformation leads to a conserved current (Noether's theorem).
- Transformations of the wave function under inversion of the space coordinates defines the parity of the system. Parity is conserved in electromagnetic and strong interactions, but not in weak interactions. As a result, fermions involved in charged-current weak interactions possess longitudinal polarization $P = \alpha\big(\sigma \cdot \mathbf{p}\big)/E = \alpha\,(v/c)$, where $\boldsymbol{\sigma}$ is the spin vector (with $\boldsymbol{\sigma}^2 = 1$), $\mathbf{p}$ and E are the three-momentum and total energy of the particle, and $\alpha = -1$ for fermions and $+1$ for antifermions.
- Particles created singly in parity-conserving interactions have to be assigned an intrinsic parity. Fermions and antifermions have opposite intrinsic parities.
- The helicity of a particle is a well-defined quantum number for ultra-relativistic particles; $H = \sigma \cdot \mathbf{p}/|\mathbf{p}| = +1$ or -1 (i.e. RH or LH). In vector or axial–vector interactions, the helicity of a relativistic particle is preserved, that is, it has the same value before and after an interaction. Neutrinos have helicity -1 and antineutrinos, helicity $+1$.
- It is believed that all successful field theories must have the property of invariance under local gauge (or phase) transformations. This leads to renormalizability of the theory, giving finite predictions to all orders in the coupling constant. In QED, local gauge invariance leads to the masslessness of the photon.
- In QED, the gauge transformation belongs to the group U(1). In the electroweak theory, the gauge transformations belong to the group SU(2) involving non-commuting operators.

- Although the mediating bosons W and Z in the electroweak theory are massive, the theory remains renormalizable as a result of the Higgs mechanism.
- In the simplest electroweak model, one physical scalar Higgs boson is predicted. The present lower limit on the mass is about 100 GeV.
- The summation over Feynman graphs of higher orders can be approximated by single boson exchange with an effective coupling which 'runs' with the momentum transfer involved.
- Taking account of the virtual processes involved (radiative corrections), the electroweak theory predicts relations between the various parameters (particle masses, mixing angles, decay asymmetries, etc.) and these have been tested experimentally to high accuracy.
- All interactions are invariant under the C, P, and T operations taken in any order. This CPT invariance results in the same mass and lifetime for particle and antiparticles and for the spin-statistics relation.
- CP invariance holds good in strong and electromagnetic interactions, but is violated in weak interactions. CP violation is allowed in the Standard Model of particle physics, with weak transformations between three families of quarks and leptons. CP violation is required to account for the matter–antimatter asymmetry of the universe, although it is not clear that the violations observed in laboratory experiments on K^0 and B-meson decays are relevant to the cosmological problem.

Problems

More challenging problems are marked with an asterisk.

(3.1) Show that, if the pions are in a state of zero relative orbital angular momentum (S-state), then $\pi^+\pi^-$ is an eigenstate of CP $= +1$ and $\pi^+\pi^-\pi^0$ is one of CP $= -1$.

(3.2) Explain why the π^+ and π^- mesons are of equal mass, while the baryons Σ^+ and Σ^-, both of strangeness $S = -1$, have masses of 1189.4 MeV/c^2 and 1197.4 MeV/c^2, respectively.

(3.3) The neutral non-strange mesons ρ^0 (spin $J = 1$, mass 770 MeV) and f^0 ($J = 2$, mass 1275 MeV) can both decay to $\pi^+\pi^-$. What are their C and P parities? State which of the decays $\rho^0 \to \pi^0\gamma$ and $f^0 \to \pi^0\gamma$ are allowed, and estimate the branching ratio.

(3.4) Show that the reaction $\pi^- + d \to n + n + \pi^0 + Q$ (where $Q = 1.1$ MeV) cannot proceed for pions at rest.

*(3.5) At energies of a few GeV, the cross-section for the electromagnetic process $e^- + p \to e^-$ + hadrons is much larger than that for the weak process $e^- + p \to \nu_e$ + hadrons. However, at high energies and at high enough values of the momentum transfer, the two processes may have comparable cross-sections. These conditions would obtain, for example, at the HERA collider at DESY, Hamburg, where 30 GeV electrons collide head-on with 820 GeV protons.

(a) Calculate the total collision energy at HERA in the overall centre-of-momentum frame of the electron and proton.
(b) If the primary collision is treated as between the electron and a quasi-free u-quark carrying 25% of the proton momentum, what is the CMS energy of the electron–quark collision?
(c) At approximately what value of the four-momentum transfer squared (q^2) between electron and quark will the electromagnetic and weak cross-sections become equal? Refer to Section 1.9 for cross-section formulae.
(d) Write down any other process of electron–proton scattering which will be important at high q^2.

*(3.6) On coming to rest in matter, a positron forms an 'atomic' S-state e^+e^- with an electron, called

positronium, which is observed to decay to two or three photons, with two distinct lifetimes.

(a) What are the quantum numbers (total angular momentum J, parity P, and charge conjugation parity C) of these states?

(b) The energy levels of the hydrogen atom are given by the formula $E_n = -\alpha^2 \mu c^2/2n^2$, where n is the principal quantum number and $\mu = mM/(m+M)$ is the reduced mass of the proton, mass M, and the electron, mass m. Calculate the $n = 2 \rightarrow n = 1$ level spacing in eV in positronium ($M = 938$ MeV/c^2, $m = 0.511$ MeV/c^2, $\alpha = 1/137$).

(c) Try to estimate the lifetimes of the two decay modes, based on the fact that electron and positron wave functions have to overlap to annihilate, and that the Bohr radius in hydrogen is $a = h/(\mu c \alpha)$.

(3.7) Electron–positron annihilation at the appropriately high ‘resonant’ energy can result in the formation of the Υ-meson (the upsilon meson) of mass 9460 MeV/c^2, which is a bound state of a ‘bottom’ quark and antiquark: $e^+e^- \rightarrow b\bar{b} \rightarrow$ hadrons. Assuming the quark pair is in a state of orbital angular momentum $L = 0$, what are the quantum numbers J^{PC} of the Υ-meson?

Energy levels due to radial excitations of the $\text{b}\bar{\text{b}}$-system are observed above the ground state, the first such level being one of mass 10,023 MeV/c^2. The corresponding $2^3S - 1^3\text{S}$ level separation in positronium is 5.1 eV (see the previous question). Estimate the value of the strong coupling α_s binding the quark and antiquark, assuming for simplicity a $1/r$ (Coulombic) interquark potential (i.e. the first term in (1.7)).

*(3.8) Write down how the following quantities will transform under the P (space inversion) and T (time reversal) operations:

Position coordinate	$\mathbf{r}$
Momentum vector	$\mathbf{p}$
Spin/angular momentum vector	$\boldsymbol{\sigma} = \mathbf{r} \times \mathbf{p}$
Electric field	$\mathbf{E} = -\nabla V$
Magnetic field	$\mathbf{B} = \mathbf{i} \times \mathbf{r}$
Electric dipole moment	$\boldsymbol{\sigma} \cdot \mathbf{E}$
Magnetic dipole moment	$\boldsymbol{\sigma} \cdot \mathbf{B}$
Longitudinal polarization	$\boldsymbol{\sigma} \cdot \mathbf{p}$

Show that an electric dipole moment for the neutron would violate T-invariance. Try to estimate an upper limit to such a dipole moment, assuming the appropriate level of CP invariance is that observed in neutral kaon decay.

Estimate the expected level of asymmetry in the scattering of polarized protons by polarized protons (the polarization being longitudinal).

(3.9) All of the following decays are allowed by energy conservation. Which of them is allowed by other conservation laws? (*Note*: The ρ-meson has $J^P = 1^-$. The π and η-mesons have $J^P = 0^-$, and their principal decay modes are to two photons):

$$\begin{array}{ll}
\rho^0 \rightarrow \pi^+ + \pi^- & \rho^0 \rightarrow \eta + \gamma \\
\rho^0 \rightarrow \pi^0 + \pi^0 & \rho^0 \rightarrow \pi^0 + \eta \\
\pi^0 \rightarrow \gamma + e^+ + e^- & \eta \rightarrow e^+ + e^-
\end{array}$$

*(3.10) In a deep inelastic neutrino–nucleon collision, the quark–parton model (Section 3.12.1) predicts a pointlike cross-section proportional to energy, as in (1.27b). Above a few GeV energy, the observed neutrino–nucleon cross-section has the value $\sigma/E = 6.7 \times 10^{-39}\,\text{cm}^2$, where E is the neutrino energy in GeV. Calculate the average fractional momentum of the nucleon which is carried by the parton in such a collision.

4 Extensions of the Standard Model

The Standard Model described in Chapter 3, which has given a magnificently accurate account of a huge range of data from accelerator experiments since its inception, does, however, have limitations. As examples, it assumes neutrinos are massless, in conflict with recent experimental results, and has theoretical difficulties with topics as diverse as the so-called hierarchy problem, or of accounting for the baryon asymmetry of the universe. As will be indicated in later chapters, it is indeed on the scale of the universe that it fails to take into account completely new forms of matter and energy, which have only really become apparent since the Standard Model was first introduced in the mid-1970s. Finally of course, it does not include gravity. Including gravity with the other fundamental interactions is still an unsolved problem. Of course, we do not know what better theory will eventually replace the Standard Model, although whatever that is, the Standard Model will surely be a part of it. In this chapter, we just outline some new directions in physics going beyond it.

4.1 Neutrinoless double beta decay

As was pointed out in Chapter 1, the question as to the nature of neutrinos—whether they are Dirac or Majorana particles—is still open. In the Standard Model, neutrinos are assumed to be massless—an assumption in accord with all the data available when the Standard Model first appeared. But, as described below, evidence from neutrino flavour oscillations, appearing after 1990 showed the neutrino masses to be finite, although very small. The fact that the masses of the light neutrinos, of order 0.1 eV/c^2, are some ten orders of magnitude less than the typical masses of the known Dirac particles, such as quarks and charged leptons, suggests that they might indeed be Majorana particles, the smallness of the mass arising from the so-called see-saw mechanism described below.

If neutrinos are indeed Majorana particles, lepton number L is violated, since neutrinos ($L = 1$ in the Dirac picture) are then identical with antineutrinos ($L = -1$). However, the proof of this can be easily found by only one type of observation, namely that of neutrinoless double beta decay.

We know that, in ordinary nuclear beta decay, the emitted electron is accompanied by an (anti)neutrino, for example, in neutron decay:

$$n \longrightarrow p + \mathrm{e}^- + \bar{\nu}_e \tag{4.1}$$

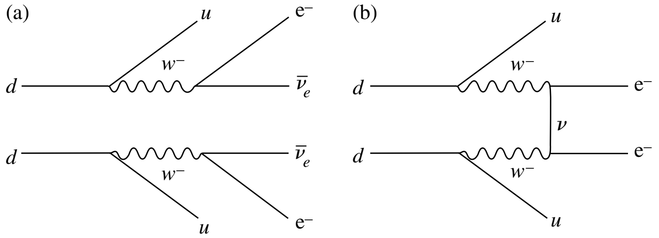


Fig. 4.1 Feynman diagrams illustrating (a) double beta decay with emission of two neutrinos; (b) neutrinoless double beta decay.

In some nuclei, two simultaneous beta decays can occur. Thus a nucleus with atomic and mass numbers (Z, A) will transform to one of $(Z + 2, A)$:

$$(Z, A) \longrightarrow (Z + 2, A) + 2\mathrm{e}^- + 2\overline{\nu}_e \tag{4.2}$$

This process is only possible for even Z, even A nuclei, because of the nuclear pairing energy. Two other conditions are necessary. First, single beta decay must be forbidden, that is, the mass of the daughter nucleus $M\,(\mathrm{Z}+1,\ \mathrm{A}) > M\,(\mathrm{Z},\mathrm{A})$, which is usually the case because of the odd–even effect on the mass. Second, energy conservation requires $M\,(\mathrm{Z}+2,\ \mathrm{A}) + 2m_e < M\ (\mathrm{Z},\mathrm{A})$. Because even–even nuclei have spin-parity 0^+, such double decays are always $0^+ \rightarrow 0^+$ transitions. These decays are second-order weak transitions, that is, the rate is proportional to G_F^4, where G_F is the Fermi constant. The measured mean lifetimes of these double beta decays are therefore very long, typically 10^{20} years or more.

If, however, neutrinos are Majorana particles, the process of *neutrinoless* double beta decay is possible:

$$(Z, A) \longrightarrow (Z + 2, A) + 2\mathrm{e}^- \tag{4.3}$$

One can think of this as a two-stage process (see Fig. 4.1). An (anti)neutrino is produced from the first decay,

$$(Z, A) \longrightarrow (Z + 1, A) + \mathrm{e}^- + \overline{\nu}_e$$

and this is absorbed by the daughter nucleus according to the equation

$$(Z + 1, A) + \nu_e \longrightarrow (Z + 2, A) + \mathrm{e}^-$$

which is of course allowed since for Majorana particles, neutrino and antineutrino are identical.

However, according to the helicity rules in weak interactions in Section 3.6, the (anti)neutrino emitted in the first process must be predominantly right-handed, while that absorbed in the second process must be predominantly left-handed. Indeed, if neutrinos were massless, such a double transition would be absolutely forbidden by helicity conservation.

For a neutrino with a finite mass m_ν and energy $E \gg m_\nu c^2$, the probability that the neutrino from the first process will emerge with the 'wrong' polarization $(1-P)$, so that it is subsequently absorbed is—see equation (3.9) and Section 3.6:

$$\left(1 - \frac{v}{c}\right) \sim \frac{(m_\nu c^2)^2}{2E^2} \tag{4.4}$$

Since in (4.3), the entire disintegration energy is shared between the two electrons, their summed energies should appear as a discrete line in the energy spectrum. (However, since it is a very rare process—if it occurs at all—the observation of neutrinoless double beta decay is likely to involve massive detectors and far from ideal resolution.) So far, neutrinoless double beta decay has not been observed, and limits on the lifetime exceed 10^{25} years. Without going into the complex details of the calculation of the nuclear matrix elements needed to compute the transition rate, and its considerable uncertainties, one can see from (4.4) that this lifetime limit must correspond to an upper limit on the neutrino mass, which is currently given in the range

$$\text{Absence of neutrinoless double beta decay} \longrightarrow m_\nu c^2 < 0.3 - 2\ \text{eV} \quad (4.5)$$

applying of course to the electron neutrino ν_e; see Klapdor-Kleingrothaus *et al.* (2001) and Fiorini (2005). This limit is still, however, considerably larger than the mass differences ($<$0.1 eV)—and by inference, the masses themselves—from observation of neutrino oscillations discussed below, and more fully described in Chapter 9. The observation of neutrinoless double beta decay would be extremely important, not simply because, if observed, it would prove beyond doubt our suspicion that neutrinos really are Majorana particles but also because from the observations one might hope to measure certain CP violating phases. Such phases could play a vital role in the decay of massive Majorana neutrinos, which could generate a lepton and baryon asymmetry in the early universe, as discussed in Section 6.5.

4.2 Neutrino masses and flavour oscillations

In the original Standard Model formulated in the 1970s, neutrinos are assumed to be massless and exist in only one (left-handed) helicity state. However, the assumed masslessness of neutrinos was questioned many years ago, in connection with the possibility of flavour oscillations. (The first proposal, by Pontecorvo, related to neutrino–antineutrino oscillations–in analogy with $K^0 - \overline{K}^0$ mixing – for which there is no evidence and which is not considered further.) Later Maki, Nakagaya, and Sakata (1962), Pontecorvo (1967), and Gribov and Pontecorvo (1969) proposed that, while neutrinos are created or destroyed as *flavour eigenstates*, they propagate through space as *mass eigenstates*. The situation is analogous to that in the quark sector, where weak interaction eigenstates are superpositions of strong flavour eigenstates. A particular neutrino flavour eigenstate, denoted by the amplitude ν_e, ν_μ, or ν_τ is therefore expressed, as regards its time evolution, as a linear superposition ν_1, ν_2, and ν_3 of mass eigenstates, which propagate through space with slightly different frequencies due to their different masses, and between which different phases develop with distance traversed, corresponding to a change or oscillation in the neutrino flavour. Thus a neutrino, created with a unique flavour, after traversing some distance in space will become a superposition of different flavours, as evidenced by any subsequent interaction with matter.

The 3×3 matrix connecting neutrino flavour and mass eigenstates is analogous to the CKM matrix (3.38) connecting quark flavour (strong interaction) eigenstates with the weak decay eigenstates. Again, the matrix involves the mixing angles between the mass eigenstates, as well as a possible CP violating phase angle. For a matrix of dimension 'n' the number of possible

(Euler) mixing angles is clearly $n(n-1)/2$, which is 3 for $n = 3$, and the possible number of non-trivial CP violating phases is $(n-1)(n-2)/2 = 1$. However, to begin with and to simplify the treatment we shall consider the case of just two flavours, with just a single mixing angle θ and no CP violation. In fact, as shown below, it turns out that, since one of the three mixing angles is very small—but from the experimentalist's point of view, hopefully not zero—and the other two are large, the actual effects observed to date can be accounted for, at least to present experimental accuracies, in terms of twofold mixing only. Using neutrino symbols to denote the wave amplitudes of the particles involved, let us, for example, consider the mixing of ν_μ and ν_τ in terms of ν_2 and ν_3 (the practical situation for atmospheric neutrinos):

$$\begin{pmatrix} \nu_\mu \\ \nu_\tau \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \nu_2 \\ \nu_3 \end{pmatrix} \tag{4.6}$$

where θ denotes some arbitrary mixing angle. The wave amplitudes

$$\begin{aligned} \nu_\mu &= \nu_2 \cos\theta + \nu_3 \sin\theta \\ \nu_\tau &= -\nu_2 \sin\theta + \nu_3 \cos\theta \end{aligned} \tag{4.7}$$

are orthonormal states. If E denotes the neutrino energy, the amplitudes of the mass eigenstates as a function of time will be

$$\begin{aligned} \nu_2(t) &= \nu_2(0)\exp(-iE_2 t) \\ \nu_3(t) &= \nu_3(0)\exp(-iE_3 t) \end{aligned} \tag{4.8}$$

where we have used units $\hbar = c = 1$, so the angular frequency $\omega = E$. The mass eigenstates will have a fixed momentum p, so that if the masses are $m_i \ll E_i$ (where $i = 2, 3$),

$$E_i = p + \frac{m_i^2}{2p} \tag{4.9}$$

Suppose that we start off at $t = 0$ with muon-type neutrinos, that is, $\nu_\mu(0) = 1$ and $\nu_\tau(0) = 0$. Inverting (4.6) we have

$$\begin{aligned} \nu_2(0) &= \nu_\mu(0)\ \cos\theta \\ \nu_3(0) &= \nu_\mu(0)\ \sin\theta \end{aligned} \tag{4.10}$$

and

$$\nu_\mu(t) = \nu_2(t)\ \cos\theta + \nu_3(t)\ \sin\theta$$

From (4.8) and (4.10) the amplitude of the muon–neutrinos becomes

$$A_\mu(t) = \frac{\nu_\mu(t)}{\nu_\mu(0)} = \cos^2\theta\ \exp(-iE_2 t) + \sin^2\theta\ \exp(-iE_3 t)$$

so that the intensity is

$$\frac{I_\mu(t)}{I_\mu(0)} = AA^* = 1 - \sin^2 2\theta\ \sin^2\left[\frac{(E_3 - E_2)\,t}{2}\right]$$

We use (4.9) and write the difference of the squares of the masses as $\Delta m^2 = m_3^2 - m_2^2$, where here and in what follows we assume $m_3 > m_2$. The probability of finding one or other flavour after a time $t = L/c$, where L is the distance travelled, is

$$P\left(\nu_\mu \longrightarrow \nu_\mu\right) = 1 - \sin^2 2\theta \cdot \sin^2\left(1.27\frac{\Delta m^2 L}{E}\right)$$
$$P\left(\nu_\mu \longrightarrow \nu_\tau\right) = 1 - P\left(\nu_\mu \longrightarrow \nu_\mu\right) \tag{4.11}$$

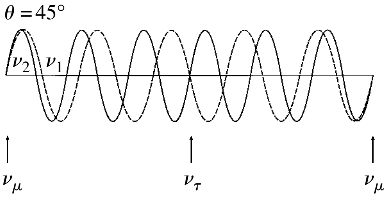


Fig. 4.2 Two neutrino ($\nu_\mu \rightarrow \nu_\tau$) oscillations, showing amplitudes of the mass eigenstates for the case $\theta = 45°$. They are in phase at the beginning and end of the plot, separated by one oscillatory wavelength, and thus from (4.6) corresponding at these points to pure muon–neutrino flavour eigenstates. In the centre of the plot the two amplitudes are 180° out of phase, corresponding to the tauon–neutrino flavour eigenstate.

Here the numerical coefficient is just $1/(4\hbar c)$ if we retain all the factors of $\hbar$ and c, and it equals 1.27 if L is expressed in km, Δm^2 in (eV/c^2)^2, and E in GeV. Figure 4.2 shows how the flavour amplitudes would oscillate for the particular case of maximum mixing, that is, $\theta = 45°$. The oscillation wavelength is $\lambda = 4\pi E/\Delta m^2$. For example, for a value of $\Delta m^2 = 3 \times 10^{-3}$ eV^2 found from the atmospheric data discussed in Section 9.15, $\lambda = 2400$ km for $E = 2$ GeV. This very long wavelength is due to the smallness of the mass difference.

In the case of three rather than two flavours, there will be three mass eigenstates m_1, m_2, and m_3 (in ascending order), with two independent mass differences, say Δm_{12} and Δm_{23} (with $\Delta m_{13} = \Delta m_{23} + \Delta m_{12}$), and three mixing angles, denoted θ_{12}, θ_{23}, and θ_{13}. Note that the oscillations only measure the *differences* of the squares of the masses rather than the masses themselves. As described in Sections 9.15–9.17, the atmospheric neutrino data and those from accelerator neutrino experiments are concerned with the larger mass difference, denoted by $|(\Delta m_{23})^2| = (m_3)^2 - (m_2)^2$ and effectively the single angle θ_{23} (denoted by Δm^2 and θ in the foregoing equations). The solar neutrino and reactor antineutrino data concern the smaller one, $|(\Delta m_{21})^2|$ and the angle θ_{12}, as shown in equation (4.12) and in Fig. 4.3:

$$\begin{array}{ll} \nu_3 \ \underline{\hspace{8em}} & \\ & \leftarrow |\Delta m_{23}|^2 \ \text{(atm)} = (2.3 \pm 0.2) \times 10^{-3}\ \text{eV}^2 \\ & \quad \tan^2\theta_{23} \qquad\quad = 1.00 \pm 0.30 \end{array} \tag{4.12a}$$

$$\begin{array}{ll} \nu_2 \ \underline{\hspace{8em}} & \\ \nu_1 \ \underline{\hspace{8em}} & \leftarrow |\Delta m_{12}|^2 \ \text{(solar)} = (8.2 \pm 0.3) \times 10^{-5}\ \text{eV}^2 \\ & \quad \tan^2\theta_{12} \qquad\quad = 0.39 \pm 0.05 \end{array} \tag{4.12b}$$

The third mixing angle has so far only an upper limit, $\tan^2\theta_{13} < 0.05$ from a reactor experiment. The signs of the above mass differences are unknown, so the mass hierarchy could be the inverse of that shown, with ν_3 being the lightest rather than the heaviest state.

The general 3 × 3 mixing matrix involving three flavours and three masses is somewhat clumsy, and can be more easily expressed as the product of three simpler matrices as follows (with $c_{23} = \cos\theta_{23}$, $s_{23} = \sin\theta_{23}$, etc):

$$U = \underset{\substack{\uparrow \\ \text{atmospheric}}}{\begin{vmatrix} 1 & 0 & 0 \\ 0 & c_{23} & s_{23} \\ 0 & -s_{23} & c_{23} \end{vmatrix}} \begin{vmatrix} c_{13} & 0 & s_{13}e^{i\delta} \\ 0 & 1 & 0 \\ -s_{13}e^{-i\delta} & 0 & c_{13} \end{vmatrix} \underset{\substack{\uparrow \\ \text{solar}}}{\begin{vmatrix} c_{12} & s_{12} & 0 \\ -s_{12} & c_{12} & 0 \\ 0 & 0 & 1 \end{vmatrix}} \tag{4.13}$$

Experimentally, it is found that θ_{13} is very small, so that $s_{13} \sim 0$ and $c_{13} \sim 1$. In this case, the middle matrix simply has unit value. The first matrix—relevant to

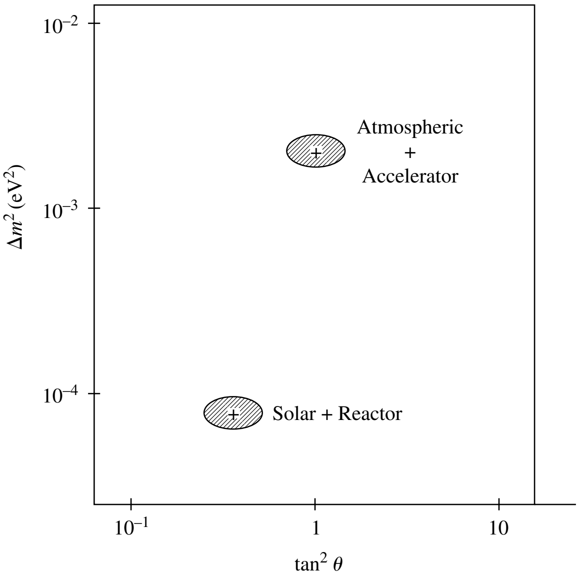


Fig. 4.3 Plots of Δm^2, the square of the mass difference, against $\tan^2\theta$ where θ is the mixing angle. At top, the combined atmospheric and long baseline accelerator data; and at bottom the solar (ν_e) and reactor ($\overline{\nu}_\mathrm{e}$) combined data (assuming CPT invariance). The boundaries of the shaded areas correspond to 90% confidence limits.

atmospheric neutrino oscillations—depends only on θ_{23}, while the third matrix, relevant to solar neutrinos, depends only on θ_{12}. The solar and atmospheric neutrinos are therefore effectively decoupled, since they are only connected *via* θ_{13}. The CP violating phase δ is seen to enter only multiplied by the small quantity $\sin\theta_{13}$, so that it will be very hard to detect. The experimental situation leading to the foregoing results is described in Chapter 9.

4.3 Grand unified theories: proton decay

The success of the electroweak theory, unifying the electromagnetic and weak interactions, in describing an enormous range of experimental data, opened the possibility that unification of the fundamental interactions might be taken one stage further, by incorporating the strong interactions with the electroweak, in what are called *grand unified theories*—GUTs for short. The basic idea is that the SU(2) × U(1) electroweak symmetry (a broken symmetry at low energies) plus the (exact) SU(3) colour symmetry of the strong interactions might be encompassed by a more global symmetry, manifested at some high unification energy, where the component symmetries would become exact. Since the effective couplings for the different interactions 'run' in different ways, the possibility arose that they might extrapolate to a universal value, the grand unified coupling α_u. This possible extension of the Standard Model was first discussed in the early 1970s, shortly after the success of the electroweak theory.

The first and simplest GUT model was the SU(5) model of Georgi and Glashow (1974). This incorporated the fermions, both leptons and quarks, into multiplets, inside which, with a common coupling, leptons and quarks could transform into one another via the exchange of massive 'leptoquark' bosons

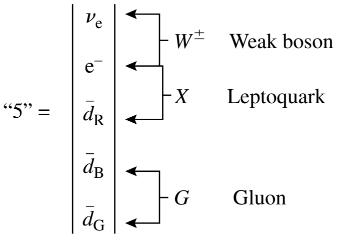


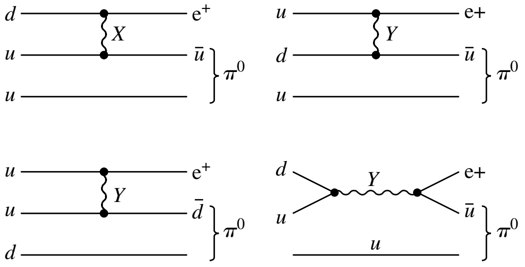
Fig. 4.4 SU(5) multiplet of quarks and leptons.

X and Y, with electric charges of 4/3 and 1/3 of the elementary charge. The diagram of Fig. 4.4 shows components of a '5' representation of SU(5), with a gluon G interacting between the quarks, the weak W boson mediating the interaction between neutral and charged leptons, and the X 'leptoquark' boson interacting between a quark and a lepton. The total charge of the multiplet is zero, corresponding to the fact that electric charge is one of the generators of the SU(5) group. Briefly, some of the attractive features of this model are as follows:

- The fractional charges of the quarks occur because the quarks come in three colours while the leptons are colourless, and the total electric charge of the multiplets is zero.
- The equality of the electron and proton charges—a historic puzzle—is accounted for.
- Because the electric charge becomes a generator of the non-Abelian SU(5) group, the commutation relations of this symmetry allow only discrete, rather than continuous eigenvalues for the electric charge. Charge quantization is thus a result of grand unification.

The unification energy, as indicated in Fig. 4.6, is in the region of 10^{14} GeV. Here the strong assumption has to be made that no other 'new' physics will enter between the electroweak scale, of order 100 GeV, and the GUT scale some 12 orders of magnitude larger—a vast energy range which has become known as the *desert*. Although the predicted value of the unification energy is far beyond reach in the laboratory, even at low energies *virtual* X and Y exchange can take place, and this would lead to the dramatic prediction of *proton decay*, for example, in the mode $p \to e^+ + \pi^0$, as indicated by the diagram in Fig. 4.5. The fact that, according to our present ideas, the observed vast asymmetry between protons and antiprotons in our world must have arisen from specific (and largely unknown) interactions in the very early universe implies, from the principle of detailed balance, that protons should decay eventually, restoring the status quo. As indicated in Fig. 4.5, because of the strong suppression factor due to the X, Y propagators, the predicted lifetime is very long, $10^{(30\pm0.5)}$ years (see Example 4.1). This is in definite contradiction with the experimental lower limit to the lifetime, which exceeds 10^{33} years. (see Fig. 4.7 for detector used to search for proton decay).

Fig. 4.5 Feynman diagrams illustrating proton decay in the SU(5) grand unification scheme. The expected lifetime is estimated in Example 4.1 and is of order 10^{30} years, that is, about one decay per day per kiloton of material. This should be easily detectable using a multikiloton detector placed deep underground to reduce cosmic ray background, such as that in Fig. 4.7. Several experiments have failed so far to find such decays, and the lower limit on the lifetime is about 10^{33} years. However, in the early 1990s these detectors found the first, and totally unexpected, evidence for flavour oscillations of atmospheric neutrinos, as discussed in Section 4.2 above and in Chapter 9.



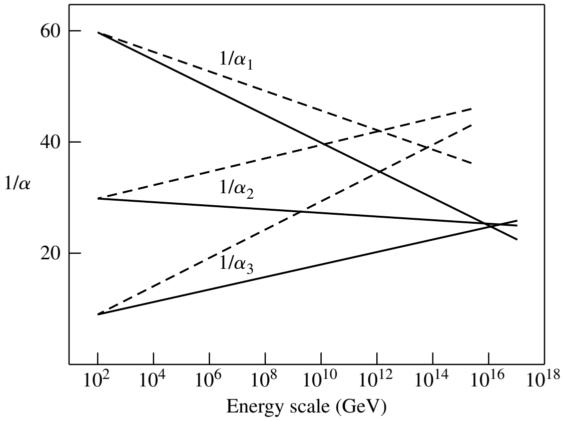


Fig. 4.6 The reciprocal couplings of the strong, electromagnetic, and weak interactions extrapolated to high energies. The dashed lines are the predictions of non-supersymmetric SU(5), and the solid lines, those predicted from supersymmetric GUT.

The three couplings involved in the grand unified model are denoted $\alpha_i = g_i^2/4\pi$, where $i = 1{-}3$. Here $g_1 = (5/3)^{1/2}\, e/\cos\theta_w$ and $g_2 = e/\sin\theta_w$, where e is the electron charge and θ_w is the weak mixing angle, refer to the electroweak couplings. The strong coupling is $g_3 = g_s$ as in (1.7). It may be shown that, in this model, the electroweak parameter has the value $\sin^2\theta_W = 3/8$ at the unification scale, where the three couplings α_1, α_2, α_3 all have the same value $\alpha_u = (8/3)\,\alpha_{\text{em}}(M_X)$. So $\alpha_u = 1/42$, using the value of $\alpha_{\text{em}}\,(M_X) = 1/112$ from Exercise 3.3, for $M_X = 10^{14}$ GeV.

The dashed lines in Fig. 4.6 show how the reciprocal quantities $1/\alpha_{1,2,3}$ vary linearly with the logarithm of the energy scale, as expected from the forms (3.32) and (3.33).

Example 4.1 *If proton decay is mediated by a boson of mass $M_X = 3 \times 10^{14}$ GeV with conventional weak coupling, estimate the proton lifetime using the value of the grand unified coupling from Example 3.3.*

An estimate can be obtained from the formula

$$\tau_p = \frac{M_X^4}{A\alpha_u^2 M_p^5}$$

where $A \sim 1$ is an arbitrary parameter giving the probability of quarks in the proton being in the correct configuration for the transition $ud \to e^+ + \bar{u}$, for example. The X-boson mass enters to the fourth power because of the propagator term, and proton mass to the fifth power from dimensional arguments. The grand unified coupling $\alpha_u = (8/3)\ \alpha_{\text{em}}(M_X) = 1/42$ from Example 3.3. Inserting these numbers (and recalling that in natural units, 1 GeV^{-1} = 9.6×10^{-25} s), results in $\tau_p = 4.3 \times 10^{29}/A$ years, where $A < 1$. The accepted value of the lifetime prediction from minimal SU(5) is $10^{(30\pm0.5)}$ years.

In summary, despite its many attractive features, the difficulty with the SU(5) model is not only that it predicts the wrong value of the proton lifetime but also that the three extrapolated couplings (Fig. 4.6) do not exactly meet at a point.

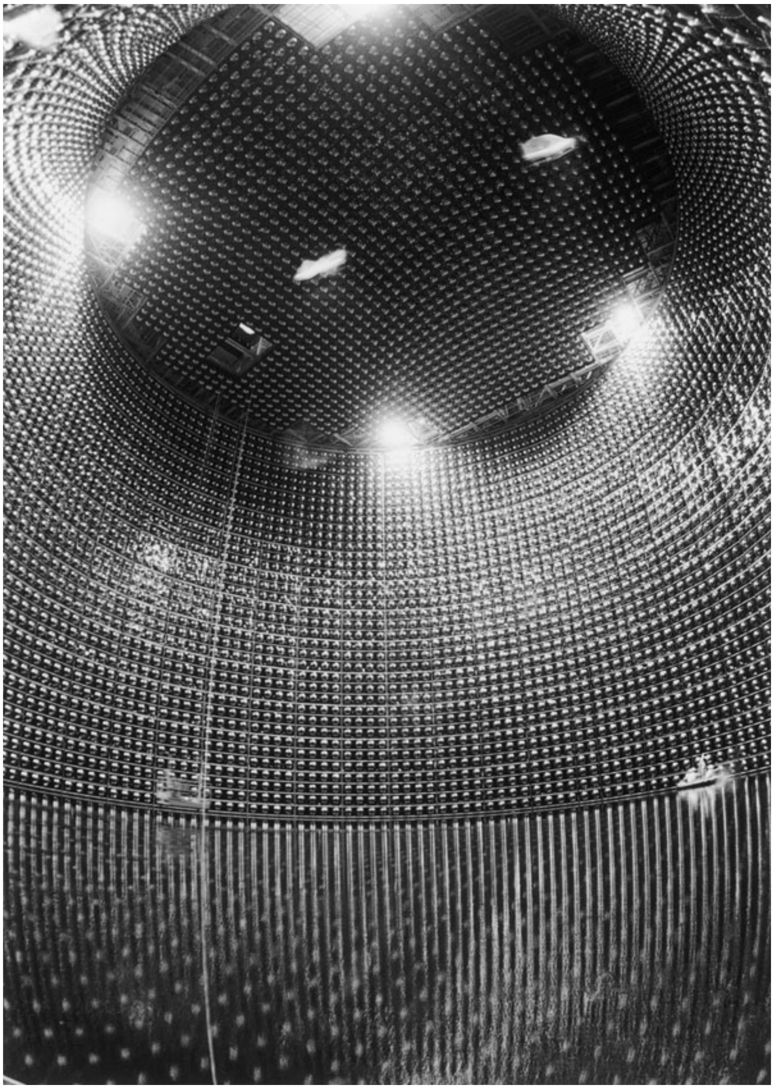


Fig. 4.7 Photograph of the Superkamiokande water Cerenkov detector used to search for proton decay. For a discussion of the Cerenkov effect, see Section 9.6. The detector consists of a cylinder of 40-m diameter and 40-m depth filled with water, with the surface covered by 11,000 photomultipliers, which record the Cerenkov light produced by relativistic charged particles as they traverse the water. This picture was taken as the tank was being filled with the full volume (50,000 tons) of water. The detector location is the Kamioka mine, Japan, at a depth of 1100 m. As will be discussed in Chapter 9, although proton decay has not yet been observed, what was thought to be an annoying background due to interactions of atmospheric neutrinos, in fact led to the very important discovery of neutrino oscillations, with the Superkamiokande detector playing a leading role. (Courtesy Prof Y. Totsuka)

4.4 Grand unification and the neutrino see-saw mechanism

A modified grand unification scheme incorporates the group called SO(10), which contains SU(5) as a subgroup. Estimates of the proton lifetime in this case are considerably longer, in the range 10^{34}–10^{38} years, and decays to strange particles, such as $p \to K + \pi$ are predicted to predominate. An important feature of the SO(10) group is that it contains a U(1) singlet state, not protected by the SU(5) gauge symmetry, so that the corresponding mass arising from radiative corrections can be very large. This is significant for the following reasons.

One of the puzzles in the neutrino sector is that, as explained in Section 4.2, neutrino mass differences, and by inference, the masses themselves, are of order 0.1 eV or less, very much smaller than the GeV-scale masses of the charged leptons and quarks. As explained in Chapter 1, charged leptons are described by the Dirac equation, and both left-handed and right-handed helicity states ψ_{L} and ψ_{R} occur. In fact, the mass term in the Dirac equation is of the form $\psi_{\mathrm{L}}\psi_{\mathrm{R}}$ (and the mass itself is obtained by multiplying by the Higgs coupling). Neutrinos, if they are pure Dirac particles (so that neutrinos and antineutrinos are distinct particle and antiparticle), would therefore have exactly zero mass if they only existed in the left-handed state ψ_{L}, as is supposed in the Standard Model. The other possibility is that they are Majorana particles, where particle and antiparticle are identical. In general, we could suppose that lepton masses result from a mixture of both Dirac and Majorana mass terms. Because, for charged leptons, particle and antiparticle are clearly distinct, they must be pure Dirac particles, while neutrinos and antineutrinos can be a mixture of the two types. Suppose we denote the Majorana masses as m_{L} and m_{R} for the left- and right-handed states. The neutrino mass matrix operating on the neutrino wavefunctions can be a combination of Majorana $m_{\mathrm{R,L}}$ and Dirac m_{D} mass terms, written in the form

$$\begin{vmatrix} m_{\mathrm{L}} & m_{\mathrm{D}} \\ m_{\mathrm{D}} & m_{\mathrm{R}} \end{vmatrix} \tag{4.14}$$

By setting this matrix into diagonal form, the eigenvalues are readily found to be

$$m_{1,2} = \frac{1}{2}\left[(m_{\mathrm{R}} + m_{\mathrm{L}}) \pm \sqrt{(m_{\mathrm{R}} - m_{\mathrm{L}})^2 + 4m_{\mathrm{D}}^2}\right] \tag{4.15}$$

Suppose now that m_{L} is very small (and we shall set it to zero), and that $m_{\mathrm{R}} = M$ is very large compared with the Dirac scale, and of the order of the GUT mass scale. Then the physical neutrino masses become

$$m_1 \approx \frac{(m_{\mathrm{D}})^2}{M} \qquad m_2 \approx M \tag{4.16}$$

One ends up therefore with a left-handed Majorana neutrino of very small mass, forced down because of the large mass M of the right-handed neutrino—what is termed the 'see-saw' mechanism. Of course, both the light- and heavy neutrinos will exist as both left- and right-handed states, and presumably will be replicated in three flavour states. If we take 10 GeV as a typical Dirac mass, a value of $M \sim 10^{12}$ Gev yields $m_1 \sim 0.1$ eV, which is in the range of light neutrino masses actually observed. This so-called see-saw mechanism therefore relies on the existence of massive Majorana neutrinos to suppress the light neutrino mass well below the Dirac mass scale.

If this view is correct, the smallness of the observed neutrino masses implies new interactions and new physics at a very-high-energy scale, perhaps that of grand unification discussed above. Furthermore, as is discussed in Chapter 6, it also appears that the decay of such massive neutrinos could result in a lepton asymmetry, converted to a baryon asymmetry by so-called instanton effects at the electroweak scale, of roughly the correct magnitude to account for the observed matter–antimatter asymmetry of the universe. So some vital features of our universe seem to hang on the knotty problem of the nature and masses of neutrinos. These predictions, arising from the results of neutrino experiments during the 1990s, of course have very deep and important implications for cosmology.

4.5 Hierarchies and supersymmetry

Another unification scheme includes the idea of *supersymmetry,* already mentioned in Section 1.3, wherein every fermion has a boson partner; conversely, for each fundamental boson there is a supersymmetric fermion partner. Supersymmetry was postulated as a way of avoiding the so-called hierarchy problem. In Chapter 3 we noted that the very successful calculation of radiative corrections to the Standard Model involved loops containing virtual fermions and bosons. However, if there exist very massive particles associated with grand unified symmetry schemes, they will be present as virtual states in such loops and lead to divergences in calculating the Standard Model parameters, unless one can arrange cancellation terms. Supersymmetry does just that, since it turns out that the amplitudes for fermion and boson loops have opposite signs (technically connected with the fact that the creation and annihilation operators for bosons and fermions obey commutation and anti-commutation relations, respectively). Thus the one-loop radiative correction to the Higgs mass is of the form $\Delta m_{\mathrm{H}}^2 \sim (\alpha/\pi)\left(m_{\mathrm{F}}^2 - m_{\mathrm{B}}^2\right)$ and, provided the mass scale of the superpartners is less than about 1 TeV, the strong (quadratic) divergences are avoided (although much smaller, logarithmic divergences remain). A bonus of this scheme is that above the SUSY (= supersymmetry) scale, the evolution of the three running couplings is modified and they do meet more nearly at a point, as shown in Fig. 4.6, with a higher unification energy of around 10^{16} GeV. Because of the larger M_X, the expected proton lifetime is longer, of order 10^{35} years, which is compatible with the experimental limit. As stated previously, in this modified unification scheme, the preferred proton decay modes are into heavier particles, for example, $p \to K^+ + \nu_\mu$.

At the present time, there is no direct experimental support for supersymmetry or for grand unification. Lower limits on the masses of SUSY particles from accelerators are ~100 GeV. This is of course larger than the masses of most of the known fundamental fermions and bosons. Clearly, supersymmetry is a broken symmetry, and it could be that all the superpartners have masses in the range 100–1000 GeV. A list of some SUSY particles is given in Table 4.1. A word should be added about notation. The boson superpartners of the fermions are denoted by adding an ‘s’ prefix; thus squark, slepton, and so on. The fermion superpartners of the bosons are denoted by adding ‘ino’ after the name; for example, photino, zino, gluino, and so on.

Most supersymmetric models postulate an R-symmetry, that is, the SUSY particles are produced in pairs with conserved quantum numbers $R = \pm 1$, in much the same way that strange particles are pair-produced with $S = \pm 1$ in conventional strong interactions. Thus a quark and antiquark with sufficient energy could annihilate to a squark–antisquark pair. A massive SUSY particle would decay, in an R-conserving cascade process, to lighter SUSY particles, and eventually to the lightest superparticle, which, in the limit of exact R-conservation, would be completely stable. If this were a photino, for example, its production from squark decay, $\widetilde{Q} \to Q + \widetilde{\gamma}$ would be manifest by acoplanarity of the decay and momentum imbalance from the missing photino.

The so-called minimal supersymmetric standard model (MSSM) of course encompasses a host of new particles, including a richer Higgs sector with five physical Higgs bosons. The SUSY partners of the four electroweak gauge bosons consist of four neutral fermions, referred to as *neutralinos*. As will be

Table 4.1 Examples of supersymmetric particles (with spin in units $h/2\pi$)

Particle	Symbol	Spin	SUSY partner	Symbol	Spin
Quark	Q	$\frac{1}{2}$	Squark	$\widetilde{Q}$	0
Lepton	l	$\frac{1}{2}$	Slepton	$\widetilde{l}$	0
Gluon	G	1	Gluino	$\widetilde{G}$	$\frac{1}{2}$
Photon	γ	1	Photino	$\widetilde{\gamma}$	$\frac{1}{2}$
Z boson	Z	1	Zino	$\widetilde{Z}$	$\frac{1}{2}$
W boson	W	1	Wino	$\widetilde{W}$	$\frac{1}{2}$
Higgs	H	0	Higgsino	$\widetilde{H}$	$\frac{1}{2}$
Graviton	g	2	Gravitino	$\widetilde{g}$	$\frac{3}{2}$

discussed in Chapter 7, one of the major problems in our understanding of the universe is to account for the nature of the 'dark matter', allegedly accounting for over 80% of the total mass. If the dark matter is in the form of elementary particles, then neutralinos, created in the primordial universe and with masses in the TeV range, are possible constituents. Because of conservation of R-parity, the lightest of the supersymmetric particles (LSP) would be stable, and is regarded as a strong candidate for dark matter. Note that, despite its large mass (at least 100 GeV) this LSP must be stable enough to survive the 14 billion years age of the universe. That is quite acceptable. After all the proton, with no absolute conservation law (gauge principle) to guarantee its stability, has a lifetime at least 10^{23} times as long as this.

There are several free parameters in supersymmetric theories, so that the predictions on some physical quantities can vary considerably. For example, the very precise limit of $< 10^{-25}|e|$ cm on the electric dipole moment of the neutron already limits the range of such parameters. As indicated in Chapter 7, if dark matter does consist of supersymmetric particles, then present experimental limits are already restrictive on the parameter ranges.

4.6 Summary

- The question of the nature of neutrinos—whether they are Dirac or Majorana particles—is still open, but the observation of neutrinoless double beta decay would prove them to be Majorana particles.
- The smallness of light neutrino masses, less than 0.1 eV/c^2, may be due to the existence of very massive Majorana neutrinos and the 'see-saw' mixing mechanism.
- Neutrinos of a particular flavour consist of a superposition of mass eigenstates. The larger of the two mass differences is associated with atmospheric neutrino oscillations, the smaller with solar neutrino oscillations. These two phenomena are effectively decoupled because one of the three mixing angles is very small and as yet not determined, the other two being large.

- The success of the electroweak model led to speculations that strong as well as electroweak interactions might be unified, with a single coupling, at some high grand unification energy scale. Such a theory would predict proton decay, and provide an understanding of the equality of electron and proton charges, the fractional quark charges, and the discreteness of the electric charges of all known particles. The original SU(5) model predicted a lifetime of 10^{30} years, in contrast with the present experimental lower limit of 10^{33} years. Supersymmetric versions of the model predict longer lifetimes, consistent with experiment, and the extrapolation of the three couplings to a single value is more convincing.
- The so-called hierarchy problem led to the postulate of supersymmetry, namely that all fermions (bosons) will have boson (fermion) partners, since the radiative corrections from bosons and fermions largely cancel. The experimental mass limits on SUSY particles are above 100 GeV. At present there is no direct evidence, either for GUTs or for supersymmetry, but SUSY particles have been postulated as candidates for dark matter.

Problems

(4.1) The flux of relativistic cosmic ray muons at sea-level is approximately 250 $\mathrm{m}^{-2}\mathrm{s}^{-1}$. Their rate of ionization energy loss as they traverse matter is about 2.5 MeV $\mathrm{gm}^{-1}\,\mathrm{cm}^2$. Estimate the annual human body dose due to cosmic ray muons, in grays or rads (1 gray = 100 rads = 1 J/kg = 6.2×10^{12} MeV kg^{-1}), and compare your answer with the measured natural dose rate of 0.3 rads (which includes that from radioactivity).

If protons were to decay, a substantial part of their total mass energy (938 MeV/c^2) would appear in the form of ionizing radiation (pions, γ-rays, etc). Assume that 100 times the natural dose rate would be lethal for advanced life forms, deduce from your very existence a lower limit to the proton lifetime.

(4.2) If proton decay is mediated by a boson of mass $M_X = 3\times10^{14}$ GeV with conventional weak coupling, estimate the proton lifetime from the fact that the muon mass is 106 MeV and the mean lifetime for its weak decay is 2.2 μs.

(4.3) In an experiment using a reactor as a source of electron–antineutrinos, the observed rate of the reaction $\overline{\nu}_e + p \rightarrow \mathrm{e}^+ + n$ in a detector placed 250 m from the reactor core is found to be 0.95 ± 0.10 of that expected. If the mean effective antineutrino energy is 5 MeV, what limits would this place on a possible neutrino mass difference, assuming maximal mixing?

Part 2

The Early Universe

This page intentionally left blank

The expanding universe

5

5.1 The Hubble expansion

Everyone is familiar with the fact that the universe is populated by stars and that these stars occur in huge assemblies called galaxies. A typical galaxy such as our own Milky Way will contain of order 10^{11} stars, together with clouds of gas and dust. Various forms of galaxy are observed. One of the most common forms are the spiral galaxies, in which the older, population II stars are located in a central spherical hub, which is surrounded by a flattened structure or disc in the form of a spiral, associated with the formation of younger, population I stars moving in roughly circular orbits, and concentrated in spiral arms. Figure 5.1(a) shows a picture of the spiral galaxy M31, which is similar in structure to our own Milky Way, sketched in Fig. 5.1(b), and which, together with our nearest neighbour galaxy, the Large Magellanic Cloud, forms part of the Local Group of around 30 galaxies. Other forms of galaxy are the oval galaxies and the irregular galaxies, the latter being important as the apparent source of γ-ray bursts, which are some of the most energetic events in the universe (discussed in Chapter 9). The total number of observable galaxies is enormous, of order 10^{11}. They occur in clusters—see Fig. 5.2(a) for the Coma cluster—and superclusters separated by enormous voids as in Fig. 5.2(b) and (c). In other words, the material of the universe is not distributed at random, but there is structure on the very largest scales. Typical sizes and masses are given in Table 5.1.

For the radius of the universe in this table we have simply quoted the *Hubble length* ct_0 where $t_0 = 14$ Gyr is the age of the universe, as discussed below. In fact, on account of the Hubble expansion, the actual radius of the visible universe—the distance to the optical horizon—is larger than the Hubble length, and according to our present ideas, about a factor 3.3 times larger (see Section 5.6). Of course, there must be parts of the universe beyond our horizon, and for all we know, it could be infinite in extent. Indeed, as will be made clear in Chapter 8, at early stages the parts of the universe beyond the optical horizon must have played a crucial role in its development.

From the last line in this table we may note that the (negative) gravitational potential energy GM^2/R and the mass energy $Mc^2$of the universe are comparable at $\sim 10^{70}$ J, so that the total energy could be quite small. As indicated later, it turns out that the measured value of the curvature parameter on very large scales is consistent with it, and the total energy, being exactly zero. Of the various arbitrary numbers which are needed to describe our universe, this zero value seems to be the only natural one!

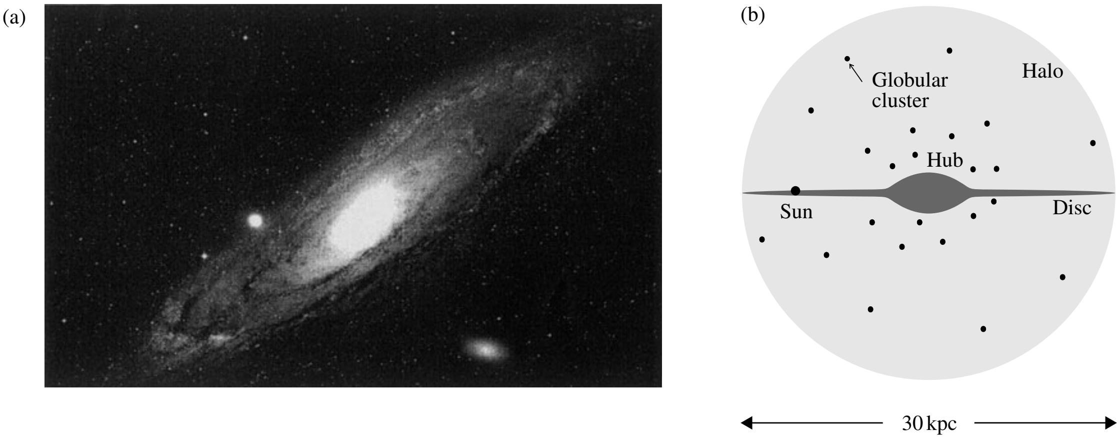


Fig. 5.1 (a) The spiral galaxy M31 in Andromeda, believed to be very similar in form to our own galaxy, the Milky Way. Two dwarf elliptical galaxies appear in the same picture. (b) Sketch of edge-on view of Milky Way. As well as stars and dust, the spiral arms of the disc contain gas clouds, predominantly of hydrogen, detected from the 21 cm wavelength emission line due to flip over of the electron spin relative to that of the proton. The Milky Way contains at least 150 globular clusters (see Fig. 10.3), each containing of the order of 10^5 very old stars of similar age. The halo region is assumed to contain dark matter as described in Chapter 7. The central hub contains a massive black hole of about 3×10^6 solar masses, identified with the X-ray/radio source Sagittarius A^* (see Section 10.11).

In 1929 Hubble, observing the spectral lines from distant galaxies with the new 100-inch Mount Wilson telescope, noted that the lines were shifted towards the red end of the spectrum, the amount of shift depending on the apparent brightness of the galaxy and hence on the distance. He measured the velocity of recession of a galaxy, v, interpreting the redshift as due to the Doppler effect (see Section 2.11). The wavelength in this case is increased from λ to λ' so that

$$\lambda' = \lambda\sqrt{\frac{(1+\beta)}{(1-\beta)}} = \lambda(1+z) \tag{5.1}$$

where $\beta = v/c$ and the redshift $z = \Delta\lambda/\lambda$. Hubble discovered a linear relation between v and the true coordinate distance D:

$$v = H_0 D \tag{5.2}$$

where H_0 is called the Hubble constant. In Hubble's early measurements, its value was vastly overestimated. As quoted in the *Particle Physics Review* (Yao *et al.* 2006)—see also the Wilkinson Microwave Anisotropy Probe (WMAP) results in Chapter 8—the usually accepted value today is

$$H_0 = 72 \pm 3 \text{ km s}^{-1} \text{ Mpc}^{-1} \tag{5.3}$$

where the megaparsec has the value 1 Mpc $= 3.09 \times 10^{19}$ km. The subscript '0' to H is to signify that this is the value measured today. In many (indeed most) texts, this number is conventionally quoted as 100h km s^{-1} Mpc^{-1} where $h =$ 0.72, because in earlier times the value of H varied widely between different observers. However, that seems hardly necessary today.

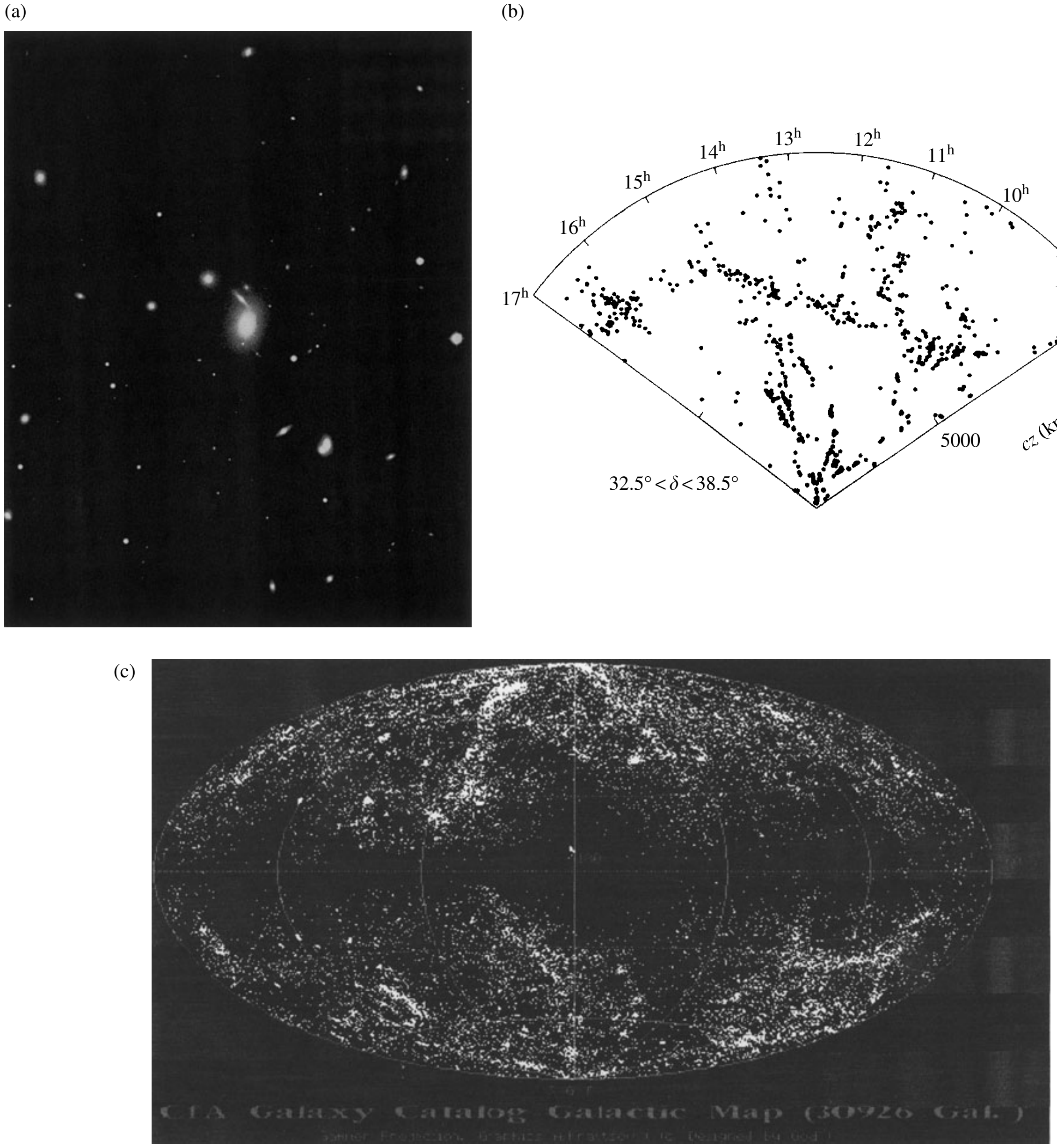


Fig. 5.2 (a) The Coma cluster of galaxies, in which both spirals and ellipticals appear. The space between galaxies in clusters is usually filled with very hot, X-ray emitting gas, which includes ions of heavy elements like iron, indicating that much of the gas is debris expelled from early generations of very massive stars which have long since disappeared from view. (Courtesy of Palomar Observatory). (b) An early plot showing the distribution of a sample of some 700 galaxies over a small range of declination angle δ. The redshift velocity cz is plotted radially, the angular coordinate being the right ascension. The existence of clusters and voids is very clear (de Lapparent *et al.* 1986). (c) Sky map of some 30,000 galaxies from the CfA catalogue, plotted in galactic coordinates. The dark horizontal band corresponds to obscuration from the plane of the Milky Way.

Table 5.1 Approximate sizes and masses in the universe. (1 parsec = 1 pc = 3.09×10^{16} m = 3.26 light years)

	Radius	Mass
Sun	7×10^8 m	2×10^{30} kg $= M_S$
Galaxy	15 kpc	$10^{11} M_S$
Cluster	5 Mpc	$10^{14} M_S$
Supercluster	50 Mpc	$10^{15} M_S$
Universe	4.2 Gpc	$10^{23} M_S$

The interpretation of the redshift in terms of the Doppler effect is permissible for the small redshifts of $z < 0.003$ observed by Hubble. For such nearby galaxies Newtonian concepts of space and time are applicable. Expanding (5.1) for small values of v/c we get

$$\lambda' \approx \lambda(1 + \beta)$$

and hence

$$z = \frac{v}{c} \tag{5.4}$$

However, for distant galaxies and large redshifts, $z \geq 1$, the Doppler formula gives $(1 + z) = \gamma(1 + \beta)$, but may not be relevant, since at such distances additional, gravitational redshifts, as described in Section 2.3, could then become important. The empirical relation observed is therefore of a linear dependence of the redshift on the distance of the galaxy, as given in the wavelength formula (5.1). The distance is estimated from the apparent brightness or luminosity, and is therefore called the *luminosity distance* D_L. It is determined from the (supposedly known) intrinsic luminosity L or total power radiated by the source (star or galaxy), and the measured energy flux F at the Earth:

$$F = \frac{L}{4\pi D_L^2} \tag{5.5}$$

In fact the astronomers use a logarithmic scale of luminosity, called *magnitude*, running (perversely) from small values of magnitude for the brightest stars to large values for the faintest. The defining relation between the apparent magnitude $m(z)$ at redshift z, the so-called absolute magnitude M (equal to the value that m would have at $D_L = 10$ pc) and the distance D_L in Mpc, is given by the *distance modulus*

$$m(z) - M = 5 \log_{10} D_L(z) + 25 \tag{5.6}$$

In the Hubble diagram, see Fig. 5.3, $(m - M)$ or $\log_{10} D_L$ is plotted against $\log_{10} z$.

A few words are appropriate here about the establishment of the ‘cosmological distance scale’. For nearby sources, distances can be measured using parallax, that is the change in direction of the source, relative to more distant sources as the Earth circulates in solar orbit. (A source at 1 parsec distance has a parallax of 1 s of arc on a baseline of the Earth–Sun distance of 1 a.u.) Over the years, a number of very clever and interlocking methods have been used to extend the distance scale. We just have space to mention here one used to measure the distance (57 kpc) to our nearest neighbour galaxy, the Large

Magellanic Cloud. This contained the supernova 1987A, of importance as it gave the first evidence for a neutrino source outside the solar system (see Section 10.9). The Hubble Space Telescope observed a ring of material, which had been ejected some 20,000 years previously when the star in question had entered the blue giant stage of evolution. This ring was seen at an angle of inclination and appeared therefore as an ellipse. Various parts of this ring appeared illuminated from the outburst at different times, because of the difference in transit times of the light to Earth. From these time differences, the inclination and the angular size of the ring, the distance could be calculated.

A modern version of the Hubble plot at small redshifts, appropriately using the Hubble Space Telescope, is shown in Fig. 5.3, for events of $z < 0.1$. The various sources in this plot include Cepheid variable 'supergiant' stars for $z < 0.01$, and Type Ia and Type II supernovae for higher redshifts. Cepheids are used as 'standard candles', since they vary in intrinsic luminosity due to oscillations of the envelope, the period τ being determined by the time for sound waves to cross the stellar material: $\tau \propto L^{0.8}$. Supernovae, discussed in Chapter 10, signal the death throes of stars in the final stages of evolution, and when they occur, their light output for a time—typically weeks or even months—can completely dominate that from the host galaxy. So in principle they are useful for probing out to large distances and redshifts, or equivalently, back to earlier times. The distance to the 30 or so nearest spiral galaxies where a few Type Ia or Type II supernovae have occurred, has been established from observations on the Cepheids, and this provides a means of calibrating supernova luminosity.

The data in Fig. 5.3 is seen to be consistent with a very constant and uniform 'Hubble flow', and this particular sample leads to $H_0 = 72$ km s^{-1} Mpc^{-1} as in (5.3). As discussed later in Section 7.14, at much higher redshifts the data indicate that H is not in fact constant with time, that it was smaller in the distant past and that the universe is now *accelerating*. However, the evidence for and implications of all this are deferred to Chapter 7.

The Hubble relation (5.2) implies a uniform and homogenous expansion of the universe with time. If H were independent of the time, it would imply an

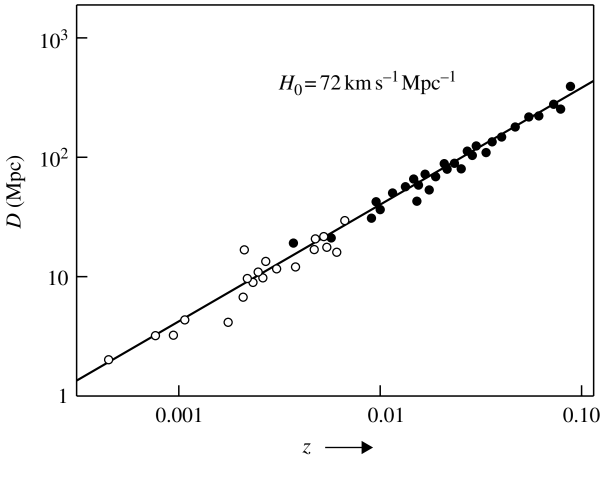


Fig. 5.3 Log–log plot of distance versus redshift, for small redshifts, $z < 0.1$. The points for $z < 0.01$ are from Cepheid variables (open circles), and those of higher z (full circles) include results from Type Ia and Type II supernovae. The straight line is that for the Hubble parameter $H_0 = 72$ km s^{-1} Mpc^{-1}. (After Freedman *et al.* 2001)

increase in the size of the universe by a factor e in the so-called Hubble time

$$t_{\mathrm{Hubble}} = \frac{1}{H_0} = 13.6 \pm 0.5 \text{ Gyr } (1.36 \times 10^{10} \text{ years}) \tag{5.7}$$

where H_0 is the current value of the Hubble parameter. The actual or *physical coordinate distance* D from the Earth, say, to some distant galaxy at time t is written as in Section 2.9 as the product

$$D(t) = r \cdot R(t) \tag{5.8}$$

where $R(t)$ is the value of the *scale parameter* and r is the *co-moving coordinate distance* measured in a reference frame which is co-moving (i.e. extending) with the expansion. The quantity r is a time-independent constant (for the distance to a particular galaxy), while according to the cosmological principle discussed later, the expansion parameter $R(t)$ is assumed to be the same over all space and depends *only* on time, in a way determined by the exact geometry (curvature) of the universe, as indicated in Fig. 5.4. Its value at time t, as compared with the value today at $t = 0$, is of course just equal to the reciprocal of the redshift factor in (5.1)

$$R(t) = \frac{R(0)}{(1 + z)} \tag{5.9}$$

One can normalize R to present-day values by defining the ratio $a(t) = R(t)/R(0)$, and in many texts the parameter $a(t)$ is used to quantify the expansion. However, in this text we will stick with R. Substituting (5.8) in (5.2) it is seen that the Hubble law is then a statement about the rate of change of the scale parameter:

$$\dot{R}(t) = H\,R(t) \tag{5.10}$$

where $\dot{R} = \mathrm{dR}/\mathrm{d}t$. The expansion can be compared with the stretching of the surface of a balloon under inflation in the two-dimensional case. However, it must be emphasized that the expansion applies *only* to truly cosmological distances, that is to those between galaxies or galaxy clusters. In the balloon analogy, the galaxies should be represented by dots or small coins of fixed diameter stuck on the balloon surface. As the balloon inflates, the galaxies remain the same size, and the pattern of the galaxies simply expands in size.

The expansion of the universe is usually referred to as the Big Bang, a nomenclature originally coined in the 1950s as a term of derision by Fred Hoyle, who was himself a devotee of the now defunct Steady State theory of the universe. How times change! The term Big Bang suggests that a sudden explosion occurred at a singular point in space–time. Obviously, referred to such an origin, this could reproduce the Hubble relation (5.2), since the particles of largest velocity will have travelled the farthest from the origin. However, the accepted view of the early universe (before formation of stars and galaxies at redshifts $z < 12$) is based on the *cosmological principle*, namely that the universe was both isotropic and homogeneous, so that no direction or location was to be preferred over any other, and thus it must appear the same to all observers no matter where they are. Observationally, the universe is indeed, even today, found to be approximately isotropic on large enough scales, and this also implies homogeneity, since a non-homogeneous universe would appear isotropic only to a favoured observer stationed at its centre, if it had spherical symmetry. So the 'Big Bangs' occur everywhere at once and the expansion is the same for all observers irrespective of their location.

Once again, we emphasize that the Hubble 'expansion of space' applies only to cosmological distances, that is on the scale of intergalactic or larger separations. It does not imply an increase with time of the size of an atom, or of a planetary system or even of a single galaxy. One can perhaps understand this on the basis of the smallness of the Hubble constant. The relation $v = HD$ in (5.2) clearly implies an outward acceleration

$$g_{\text{Hubble}} = H^2 D = 5.10^{-36} D \text{ ms}^{-2}$$

where D is in metres. It is left as an exercise to show that for the Earth–Sun system, this is only 10^{-22} of the gravitational acceleration of the Earth in solar orbit, while for a hydrogen atom, the Hubble acceleration is 80 orders of magnitude less than the inward acceleration of the electron due to the electric field of the proton. Only when we come to galactic masses $M \sim 10^{41}$ kg and intergalactic distance scales $D > 1$ Mpc do we find the (inward) gravitational acceleration $g_{\text{grav}} < 10^{-14}$ ms^{-2} exceeded by $g_{\text{Hubble}} > 10^{-13}$ ms^{-2}. Here it should also be pointed out that individual galaxies, just like individual stars, have their own 'peculiar velocities' (produced by the effects of nearby gravitating masses) relative to the general outward Hubble flow. For example, our neighbouring galaxy, M31 (see Fig. 5.1) is actually moving *towards* the Milky Way. So the Hubble expansion describes a general cosmic-scale behaviour, after peculiar velocities of individual galaxies have been averaged out.

What is the cause of the Hubble expansion? That is unknown. We simply have to accept it as an empirical fact. The reader is referred to Problem 5.6 for an early proposed model of expansion, demolished by a clever (and even 50 years later, still the best) 'table top' experiment, set up and completed within 10 days of the original proposal being made!

5.2 Olbers' paradox

In the nineteenth century, Olbers asked the question 'Why is the sky dark at night?' He supposed the universe to be unlimited in extent and filled uniformly with sources of light (stars). The light flux reaching us from a star at distance r varies as r^{-2}, while the number of stars in the spherical shell $r \to r + \mathrm{d}r$ varies as $r^2\mathrm{d}r$. Hence the total light flux will increase as r(max), which is infinite in the model.

There are several reasons why Olbers' arguments are invalid. First, we believe the observable universe is not infinite but has a finite age, and began at a time t_0 in the past with the Big Bang which started off the Hubble expansion. This means that light can only reach us from a maximum horizon distance ct_0, and the flux must be finite. A second point is that the light sources (stars) are finite in size, so that nearby sources will block out light from more distant sources. Their light is absorbed exponentially with distance by the intervening stars and dust. Third, stars only emit light for a finite time t, and the flux from the most distant stars will therefore be reduced by a factor t/t_0. Finally, the expansion of the universe results in an attenuation at large enough redshifts of light of any particular frequency; for example, red light will disappear into the infrared and the flow of light energy will fall off. However, we may remark that, as indicated below, the 2.7 K microwave background radiation which is the cooled and red-shifted remnant of the original expanding fireball of the Big Bang, although invisible to the eye, is just as intense at night time as during the day. So in this sense Olbers was right!

5.3 The Friedmann equation

The evolution of the universe can be described theoretically by the solution of Einstein's field equations of general relativity. The 'Standard Model' of present day cosmology is the solution proposed by Friedmann–Lemaitre–Robertson–Walker (FLRW for short), which assumes a completely isotropic and homogenous distribution of matter and radiation, behaving like a perfect frictionless fluid. This assumption of isotropy and homogeneity is a statement of the cosmological principle mentioned above. Of course, whatever was the case at early times in the universe, the matter today does show enormous fluctuations in density in the form of stars, galaxies and larger structures. But even today the average separation between galaxies is of order 100 times their diameter and the overall expansion of the universe of billions of galaxies on large enough scales, that is many orders of magnitude larger than the intergalactic separations, still appears to be reasonably well described by the FLRW model. Thus the universe is homogeneous in the same sense as is a volume of gas on a scale large compared with the intermolecular separation. However, the best evidence for isotropy and homogeneity on all scales in fact comes from observations of the cosmic microwave background (CMB) discussed in Section 5.7, which reflects the distribution of matter and radiation when the universe was only about 380,000 years old (as compared to 14 Gyr today) and long before either stars or larger structures had started to form.

The solution for the temporal development of the universe predicted by this model was first found by Friedmann (1922), and for the time components of the field equations has the form (see also (5.20)):

$$H^2 = \left(\frac{\dot{R}}{R}\right)^2 = \frac{8\pi G\rho_{\mathrm{tot}}}{3} - \frac{kc^2}{R^2} \tag{5.11}$$

where $R = R(t)$ is the expansion parameter in (5.8) and (5.9), ρ_{tot} is the total density of matter, radiation, and vacuum energy, as described below, and G is Newton's gravitational constant. This equation follows when the FLRW metric (2.31) is inserted in the Einstein field equations (2.19).

The term kc^2/R^2 is the *curvature term*. As indicated in Chapter 2, one of the consequences of general relativity is that, in the presence of gravitating masses, the 'flat' Euclidean space of special relativity is replaced by curved space/time. A light beam going from point A to point B will travel along a path which is an extremum, which is the shortest spatial path length (and also the path of maximum proper time) called a *geodesic*. Geodesics are straight lines in Euclidean space, but in the presence of gravitational fields, the paths are curved. Of course, this idea is an extension to three dimensions, of the familiar two-dimensional case in which the shortest path between points on a spherical surface is along a great circle.

In the language of particle physics, one might say that space/time appears curved because photons are deflected (and also retarded, as we know from Section 2.7) by gravitational fields, mediated by graviton exchange. The curvature parameter k can in principle assume values of $+1$, 0, or -1, corresponding to the curvature k/R^2 being positive, zero, or negative respectively. The two-dimensional analogy for positive or convex curvature is the surface of a sphere, while that for negative or concave curvature is as in a saddle.

Equation (5.11) has simply been quoted without derivation, but we can understand in terms of Newtonian mechanics what it implies in the special case where the energy density is dominated by non-relativistic matter. Let us consider a point mass m being accelerated by gravity at the surface of a sphere of radius D, density ρ, and mass $M = 4\pi D^3\rho/3$. According to Newtonian mechanics, the assumed spherically symmetric and homogeneous distribution of matter outside of the sphere can make no contribution to the force, while the field at the surface is the same as if all the mass were concentrated at the centre. It turns out that this is also true in general relativity (by a theorem due to Birkhoff). So the force equation is simply

$$m\ddot{D} = -\frac{mM\mathrm{G}}{D^2} \tag{5.12}$$

where $\ddot{D} = \mathrm{d}^2D/\mathrm{d}t^2$. In this equation, if we express M in terms of D and ρ, factors of r from (5.8) cancel out, and for brevity in what follows we choose units such that $r = 1$ (but must remember that all true cosmological distances are the product $R(t)r$). After integrating (5.12) we then get

$$\frac{m\dot{R}^2}{2} - \frac{mM\mathrm{G}}{R} = \text{constant} = -\frac{mkc^2}{2} \tag{5.13}$$

If we multiply through by $2/mR^2$, we obtain an equation in agreement with (5.11), after setting the constant of integration equal to the value given by general relativity. We note that the terms on the left-hand side of (5.13) correspond to the kinetic and potential energies of the mass m, and so the so-called curvature term on the right simply represents the total energy. $k = -1$ corresponds to negative curvature and positive energy, that is to say an *open* universe expanding without limit. For $k = -1$, $\dot{R}(t) = r \cdot \dot{R}(t) \to c$ at large enough values of R. A value $k = +1$ corresponds to a *closed* universe with negative total energy and positive curvature, which reaches a maximum radius and then collapses. $k = 0$ is the simplest case, where the kinetic and potential energies just balance and the total energy and curvature are both zero. The universe expands for ever but the velocity tends asymptotically to zero at large t. This case is called the *flat* universe. These three cases are illustrated in Fig. 5.4.

Present data indicate that on large scales the universe is extremely close to being flat, with $k \approx 0$. In that case the constant of integration on the right-hand side of (5.13) representing the total energy, potential plus kinetic, is practically zero.

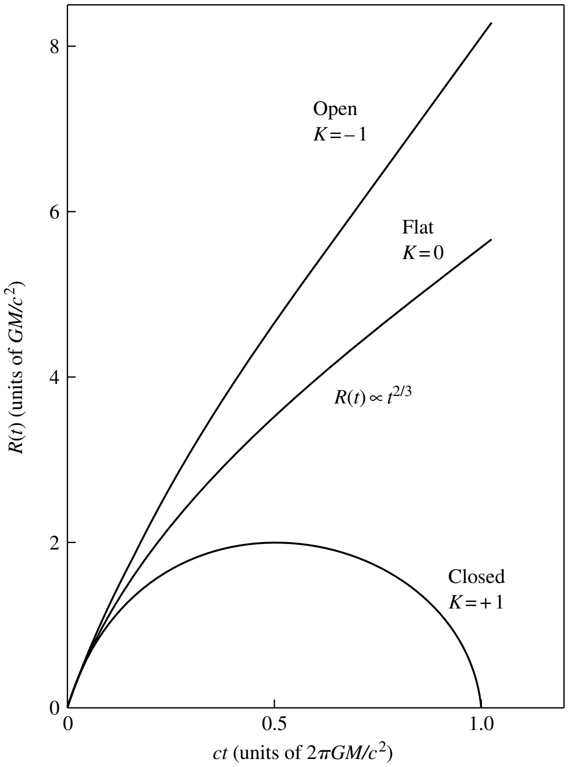


Fig. 5.4 Dependence of the scale factor $R(t)$ on time for three different k values. At the present time, our universe appears to be extremely close to the $k = 0$ curve, as discussed in more detail in Chapter 7. At early times, the parameter $R(t)$ varies as $t^{2/3}$ for all k values (see Example 5.2). For a vacuum-dominated universe, on the contrary, the scale factor increases exponentially with time (see Table 5.2). In the distant past, when the universe was only half its present age, it seems that it was indeed matter dominated. However, at the present time, the contribution to the total energy density from the vacuum is more than twice that due to matter (see Section 5.5). The present age of the universe (5.16) corresponds to the very early part of the $k = 0$ curve (roughly 0.15 on the x-axis scale).

Example 5.1 *Show that for a curvature term with* k $= +1$*, the Big Bang would be followed by a Big Crunch at time* $t = 2\pi GM/c^3$*where M is the (assumed conserved) mass of the universe.*

For $k = +1$, the Friedmann equation becomes $\left(\dot{R}/R\right)^2 = 2GM/R^3 - c^2/R^2$. From this expression it is clear that $\dot{R} = 0$ when $R = 2GM/c^2$, which is the maximum radius. The element of time is then given by $\mathrm{d}t = \mathrm{d}R/\left(2GM/R - c^2\right)^{1/2}$. Substituting $2GM/R - c^2 = c^2 \tan^2\theta$ the total time to the maximum is $\left(4GM/c^3\right) \int \cos^2\theta \mathrm{d}\theta = GM\pi/c^3$. By symmetry the time to the subsequent crunch is just twice this. The value of M from Table 5.1 gives $t \sim 100$ Gyr, corresponding to unity on the scale of the x-axis in Fig. 5.4.

Upon integrating (5.11) for the case $k = 0$ and a universe dominated by non-relativistic matter of conserved mass M one finds

$$R(t) = \left(\frac{9GM}{2}\right)^{1/3} t^{2/3} \tag{5.14}$$

so that the Hubble time (5.7) is $1/H_0 = R(0)/\dot{R}(0) = 3t_0/2$ and the age of the universe is then

$$t_0 = \frac{2}{(3H_0)} = 9.1 \pm 0.2 \text{ Gyr} \tag{5.15}$$

Other estimates of the age of the universe give significantly larger values. They come, for example, from study of the luminosity–colour relation (Herzsprung–Russell diagram) in the oldest star populations, the globular clusters (see the caption to Fig. 10.3); from cooling rates of white dwarf stars; and from dating using isotopic ratios of radioactive elements in the Earth's crust and in very old stars. All these estimates straddle an approximate range for the age of the universe of

$$t_0 = 14 \pm 1 \text{ Gyr.} \tag{5.16}$$

The discrepancy between this figure and that for a flat, matter-dominated universe (5.15) could be due in principle either to curvature ($k \neq 0$) or to the existence of a cosmological constant, as discussed below (see (5.23)). However, measurements to be described later indicate that $k \approx 0$. In fact as shown in Example 5.3, when account is taken of the effect of the vacuum energy/cosmological constant, the age of the universe estimated from the Hubble parameter is in excellent agreement with the result (5.16). Indeed, it is quite remarkable that completely independent estimates of the age come out in agreement to within 5% or so.

Example 5.2 *Find solutions of the Friedmann equation for the case of a matter-dominated universe of total mass M, and values of* $k = +1$ *and* $k = -1$.

The Friedmann equation (5.11) in this case takes the form $\dot{R}^2 = 2MG/R - kc^2$. For $k = +1$, the solution for R as a function of t has the parametric form of a cycloid curve (i.e. the curve traced out by a point on the circumference of a circular disc rolling along a plane):

$$R = a(1 - \cos\theta)$$

$$t = b(\theta - \sin\theta) \tag{5.17}$$

as can be verified by substitution. Here, the constants $a = MG/kc^2$ and $b = MG/\left(kc^2\right)^{3/2}$, and the parameter θ is the angle of rotation of the cycloid. For the case $k = -1$, the corresponding solution is

$$R = a(\cosh\theta - 1)$$

$$t = b(\sinh\theta - \theta) \tag{5.18}$$

with the above values of a and b, and k replaced by $|k|$. The curves in Fig. 5.4 were plotted from these expressions, and the solutions for the maxima and minima in Example 5.1 are found by setting $\theta = \pi$ and 2π in (5.17).

By expanding the above circular functions for small values of θ, it is straightforward to show that for either $k = +1$ or -1, the expansion parameter $R \propto t^{2/3}$, that is the same as for the case $k = 0$ in (5.14).

5.4 The sources of energy density

The conservation of energy E in a volume element $\mathrm{d}V$ of our perfect cosmic fluid can be expressed as

$$\mathrm{d}E = -P\mathrm{d}V$$

where P is the pressure. Then with ρc^2 as the energy density this becomes

$$\mathrm{d}\left(\rho c^2 R^3\right) = -P\mathrm{d}(R^3)$$

which leads to

$$\dot{\rho} = -3\left(\frac{\dot{R}}{Rc^2}\right)\left(P + \rho c^2\right) \tag{5.19}$$

Differentiating (5.11) and substituting for ρ we get the differential form of the Friedmann equation (which results from solving the spatial components of the Einstein field equations)

$$\ddot{R} = -\left(\frac{4\pi \mathrm{G}R}{3}\right)\left(\rho + \frac{3P}{c^2}\right) \tag{5.20}$$

which is the same as (5.12) for the case $P \approx 0$ for non-relativistic matter. Generally, the quantities ρ and P will be connected by an equation of state, which can be written in the general form

$$P = w\rho c^2 \tag{5.21a}$$

where w is a parameter which may be constant, as it is for matter, radiation, and the vacuum state, or might be time dependent. If w is in fact time independent, then from (5.19) and (5.21a) one obtains the simple relation

$$\rho \propto R^{-3(1+w)} \tag{5.21b}$$

as can be verified by substitution. The variation of ρ with R follows for the different regimes shown in Table 5.2.

The overall density ρ in the Friedmann equation is generally considered to be made up of (at least) three components, corresponding to the contributions from matter, radiation, and the vacuum state:

$$\rho_{\mathrm{tot}} = \rho_m + \rho_r + \rho_\Lambda \tag{5.22}$$

The quantity ρ_Λ, which we have here identified with the vacuum state, can be incorporated in the Friedmann equation as a *cosmological constant* Λ, such that

$$\rho_\Lambda = \frac{\Lambda}{8\pi \mathrm{G}} \tag{5.23}$$

The quantity Λ had originally been introduced by Einstein, before the advent of the Big Bang hypothesis, in an attempt to achieve a static (non-expanding and non-contracting) universe. Clearly, if a term $\Lambda/3$ is added to the right-hand side of (5.11), then at large enough $R(t)$ this term will dominate and the expansion will become *exponential*, that is, $R(t) \propto \exp(\alpha t)$ where $\alpha = (\Lambda/3)^{1/2}$. Present evidence, discussed in Chapters 7 and 8, indicates a finite value of Λ, with ρ_Λ larger than ρ_m (see below).

In Table 5.2, the dependencies of ρ on $R(t)$ and of $R(t)$ on t are given for different possible regimes, namely a radiation-dominated, matter-dominated, or a vacuum-dominated universe. The equations of state for radiation and for non-relativistic matter are found as follows. Suppose we have an ideal gas consisting of particles of mass m, velocity v, and momentum mv, confined within a cubical box of side L with walls with which the particle collides elastically (see

Table 5.2 Energy density and scale parameters for different regimes

Dominant regime	Equation of state	Energy density	Scale parameter
Radiation	$P = \frac{\rho c^2}{3}$	$\rho \propto R^{-4} \propto t^{-2}$	$R \propto t^{1/2}$
Matter	$P = \left(\frac{2}{3}\right) \rho c^2 \times \left(\frac{v^2}{c^2}\right)$	$\rho \propto R^{-3} \propto t^{-2}$	$R \propto t^{2/3}$
Vacuum	$P = -\rho c^2$	$\rho = \text{constant}$	$R \propto \exp(\alpha t)$

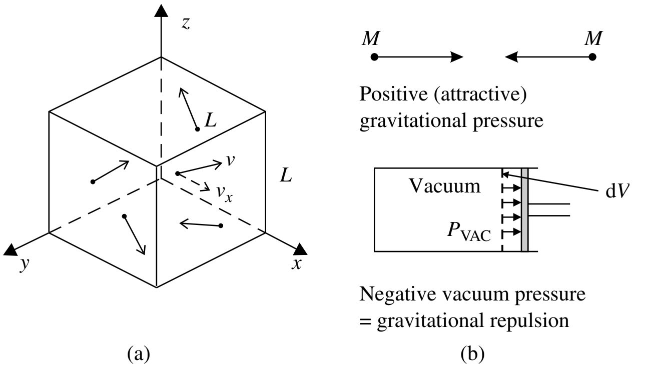


Fig. 5.5

Fig. 5.5(a)). A particle with x-component of velocity v_x will strike a particular face normal to the x-axis at a rate of $v_x/2L$ collisions per unit time. As the component of momentum $p_x = mv_x$ is reversed at each collision, the rate of change of momentum and therefore the force exerted by the particle will be $2mv_x \cdot (v_x/2L)$. The pressure exerted by the particle on the face of the box, which has area $A = L^2$, is therefore mv_x^2/L^3, where $V = L^3$ is the volume. If there are n particles per unit volume, it follows that the pressure they exert will be $mn\langle v_x^2 \rangle$ where $\langle v_x^2 \rangle$ is a mean square value. Since the gas is isotropic, the mean square values of the x, y, and z components of velocity will be equal and the pressure will be

$$P = \left(\frac{1}{3}\right) mn \left\langle v^2 \right\rangle = \left(\frac{1}{3}\right) n \left\langle pv \right\rangle \tag{5.24a}$$

Let us first assume that the gas consists of *non-relativistic* particles. Then the values of the kinetic energy density ε and of the pressure are

$$\varepsilon = \left(\frac{1}{2}\right) mn \langle v^2 \rangle$$
$$P_{\text{non-rel}} = \left(\frac{2}{3}\right) \varepsilon = \left(\frac{2}{3}\right) \rho c^2 \times \left(\frac{v^2}{c^2}\right) \tag{5.24b}$$

where ρc^2 is the total energy density of matter, including the mass energy. Since for cosmic matter in general, $v^2 \ll c^2$, the pressure it exerts is very small.

If the gas particles have *extreme relativistic* velocities, then the energy density and the pressure, usually called the *radiation pressure*, have the values

$$\rho c^2 = n\,mc^2 = n\langle pc\rangle$$

$$P_{\text{rel}} = \frac{\rho c^2}{3} \tag{5.25}$$

That the vacuum may contain an energy density and exert a pressure equivalent to a *gravitational repulsion* may seem strange, since in classical physics, a vacuum is supposed to contain absolutely nothing. However, in quantum field theory, as has been discussed for the electroweak model in Chapter 3, the Uncertainty Principle actually requires that the vacuum contains virtual particle–antiparticle pairs which spontaneously appear and disappear, and the vacuum itself is defined, not as nothing but as the state of lowest possible energy of the system. Because the virtual particles carry energy and momentum, if only on a temporary basis, general relativity implies that they must be coupled to gravitation. Indeed, the measurable effect of such vacuum energy is through its gravitational influence.

The relation $P = -\rho \mathrm{c}^2$ for this lowest energy vacuum state can be formally shown to follow from Lorentz invariance, that is, the requirement that the state must look the same to all observers, implying also that the energy density must have the same constant value everywhere and for all time. A plausibility argument for the pressure–density relation is as follows. Assume that we have a piston enclosing an isolated cylinder filled with the vacuum state of energy density ρc^2 (see Fig. 5.5(b)). If the piston is withdrawn adiabatically by an element of volume dV, the extra vacuum energy created will be ρc^2 dV, and this must be supplied by the work done by the vacuum pressure, PdV. Hence by energy conservation $P = -\rho c^2$ and from (5.19), $\rho =$ constant.

Note that in (5.20), the deceleration – $\ddot{R}$ is due to the gravitational attraction associated with the density ρ *plus* the pressure P. An increase in pressure due to relativistic particles is proportional to an increase in their energy density and hence in their gravitational potential, through the Einstein relation $E = mc^2$. Thus a negative pressure will correspond to a gravitational repulsion and the exponential expansion indicated in Table 5.2.

5.5 Observed energy densities

For the case $k = 0$, (5.11) gives a value for the *critical density* which (today) would just close the universe:

$$\rho_{\mathrm{c}} = \left[\frac{3}{(8\pi G)}\right] H_0^2 = 9.6 \times 10^{-27}\ \ \mathrm{kg\,m^{-3}}$$

and

$$\rho_{\mathrm{c}} c^2 = 5.4 \pm 0.5\ \mathrm{GeV\,m^{-3}} \tag{5.26}$$

taking H_0 from (5.3). In the second line, we have quoted a critical energy density, $\rho_c c^2$. The ratio of the actual density to the critical density is called the *closure parameter* Ω, which at the present time from (5.11) and for arbitrary k is

$$\Omega = \frac{\rho}{\rho_c} = 1 + \frac{kc^2}{[H_0 R(0)]^2} \tag{5.27}$$

One sees that a flat universe with $k = 0$ will have $\Omega = 1$ for all values of t. The different contributions to the total value of Ω are then, in parallel with (5.22) for radiation, non-relativistic matter, and vacuum densities respectively

$$\Omega = \Omega_r + \Omega_m + \Omega_\Lambda \tag{5.28}$$

If $k \neq 0$, one can express the curvature term as

$$\Omega_k = \frac{\rho_k}{\rho_\mathrm{c}} = -\frac{kc^2}{[H_0 R(0)]^2} \tag{5.29}$$

when from (5.27) one obtains

$$\Omega + \Omega_k = \Omega_r + \Omega_m + \Omega_\Lambda + \Omega_k = 1 \tag{5.30}$$

At the present time, as described in the next section, the density of the microwave photon radiation corresponds to $\Omega_r = 5 \times 10^{-5}$ as in (5.54), and is completely negligible in comparison with that of matter, while as indicated below, the vacuum term Ω_Λ makes a major contribution. In addition to the microwave photon relics of the Big Bang, there are relic neutrinos and antineutrinos (discussed in the following sections), with comparable number density and quantum energy to the photons. Because these neutrinos have masses comparable with or larger than their kinetic energies today, they are non-relativistic. However, at early times in the universe when it was much hotter and radiation-dominated, they were extreme relativistic and so would be included in the radiation term, increasing the radiation energy density by some 58%.

We now anticipate later results to be described more fully in Chapter 8. Apart from a very small energy density contribution in microwave photons and neutrinos discussed in Sections 5.8–5.11, the energy density of the universe today is made up of several components as follows:

1. For *luminous baryonic matter* (i.e. visible protons, neutrons, and nuclei) in the form of stars, gas, and dust it is found that

   $$\rho_\mathrm{lum} = 9 \times 10^{-29} \text{ kg m}^{-3}$$

   or

   $$\Omega_\mathrm{lum} = 0.01 \tag{5.31}$$

2. The *total density of baryons*, visible or invisible, as inferred from the model of nucleosynthesis described in the next chapter, is about 0.26 baryons m^{-3}, or an energy density

   $$\rho_b = 4.0 \times 10^{-28} \text{ kg m}^{-3}$$

   and

   $$\Omega_b = 0.042 \pm 0.004 \tag{5.32}$$

3. The *total matter density*, as inferred from the gravitational potential energy deduced from galactic rotation curves (see Section 7.2) and the kinematics of large-scale structures in the universe (see Section 8.9) is found to be
$$\rho_m = 2.2 \times 10^{-27}\ \mathrm{kg\,m^{-3}}$$
and
$$\Omega_m = 0.24 \pm 0.03 \tag{5.33}$$
4. The *dark (or vacuum) energy density* can be measured from an observed curvature in the Hubble plot, obtained from study of Type 1A supernovae at large redshifts (see Section 7.14). It may also be inferred from the position of the 'acoustic peaks' in the angular power spectrum of the temperature fluctuations in the microwave radiation (see Sections 8.13–8.16) measuring the total density
$$\Omega_{\text{total}} = 1.0 \pm 0.02 \tag{5.34}$$

These results and (5.33) indicate a value for the dark energy density of

$$\Omega_\Lambda = 0.76 \pm 0.05 \tag{5.35}$$

We should note here that in many texts, H_0 is specified as $100h$ km s^{-1} Mpc^{-1} where $h = 0.72$. In that case the critical density in (5.26) would be quoted as ρ_c/h^2 and the value of the closure parameter as Ωh^2, where $h^2 = 0.52$.

There are several important conclusions from equations (5.28) to (5.35). First, the value of unity in (5.34) for the total closure parameter indicates a flat universe ($k = 0$), as is predicted by the inflationary model of the very early universe described in Chapter 8. Next, we note that most of the baryonic matter is non-luminous, and that baryons, visible or invisible, account for only a small fraction, of order 17%, of the total matter. The bulk is ascribed to *dark matter*, as discussed in detail in Chapter 7. The nature of such dark matter is presently unknown. Finally, it appears that at the present epoch, the bulk of the energy density is in the form of *dark energy*. Here, as in Chapter 7, we have identified this dark energy with vacuum energy and a 'Λ' subscript, but other possibilities have been proposed, such as a fifth type of fundamental interaction, with the dark energy density being a function of time. Like dark matter, the source of the dark energy is unknown at the present time. In fact it has been possible to measure the parameter $w = P/\rho c^2$ in the equation of state (5.21) for the dark energy term, with the present result

$$w_{\text{(dark energy)}} = -0.97 \pm 0.08 \tag{5.36}$$

consistent with the value -1 for a simple vacuum. However, the crucial and important fact to bear in mind here is that *at the present time, the nature of 95% of the energy density of the universe is completely unknown.*

Other possibilities, instead of dark matter and/or dark energy, are of deviations from Newton's inverse square law of gravitation for very large cosmic distances. Such deviations from conventional gravity have been repeatedly proposed over the years, but at present there seems to be absolutely no evidence in their favour. Indeed, instances have been found of galaxies passing through one another, in which visible matter and dark matter (detected from its gravitational influence *via* gravitational lensing) are clearly separated, presumably because the visible matter undergoes electromagnetic interactions, while the dark matter is only weakly interacting. It is also perhaps worth stating here that as a result of recent measurements, the postulated partition of energy among the various components described above has changed dramatically over the last decade or so. Twenty years ago, it was thought that vacuum energy would make only a minor contribution, and that the value $\Omega \sim 1$ would be made up largely of dark matter.

Finally, we may note that it appears somewhat miraculous that, of all the conceivable values of Ω_{tot}, the one observed today appears to be very close to the value of unity expected for a flat universe with zero total energy and zero curvature. As stated before, of all the different numbers required to describe the universe, this zero value appears to be the only natural one.

5.6 The age and size of the universe

An estimate of the *age of the universe*, including all the sources of energy density, can be made as follows. From (5.11) and (5.27) to (5.30) the Hubble parameter at time t is given by

$$\begin{aligned} H(t)^2 &= (8\pi G/3)\,[\rho_m(t) + \rho_r(t) + \rho_\Lambda(t) + \rho_k(t)] \\ &= H_0^2\,[\Omega_m(t) + \Omega_r(t) + \Omega_\Lambda(t) + \Omega_k(t)] \\ &= H_0^2\left[\Omega_m(0)(1+z)^3 + \Omega_r(0)(1+z)^4 + \Omega_\Lambda(0) + \Omega_k(0)(1+z)^2\right] \end{aligned} \tag{5.37}$$

where we have used the fact that $R(0)/R(t) = (1+z)$ from (5.9), and that, as shown in Section 5.8 and Table 5.2, matter, radiation, and curvature terms vary as $1/R^3$, $1/R^4$, and $1/R^2$ respectively. The vacuum energy, by definition, is independent of z, while $\Omega_k(0) = -kc^2/(R_0H_0)^2$ as in (5.29). Furthermore, from (5.9)

$$H = \left(\frac{1}{R}\right)\frac{\mathrm{d}R}{\mathrm{d}t} = -\left(\frac{\mathrm{d}z/\mathrm{d}t}{(1+z)}\right)$$

and hence

$$\mathrm{d}t = -\frac{\mathrm{d}z}{[(1+z)H]} \tag{5.38}$$

We integrate to obtain the interval from the time t when the redshift was z, to the present time, t_0, when $z = 0$:

$$t_0 - t = \frac{1}{H_0}\int \frac{\mathrm{d}z}{\begin{array}{c}(1+z)\left[\Omega_m(0)(1+z)^3 + \Omega_r(1+z)^4\right.\\ \left.+\Omega_\Lambda(0) + \Omega_k(0)(1+z)^2\right]^{1/2}\end{array}} \tag{5.39}$$

The age is found by setting the upper limit as $z = \infty$ at $t = 0$. In the general case this integral has to be evaluated numerically, but there are a few cases where analytical solutions are possible, for example, when the radiation term can be neglected and either $\Omega_\Lambda = 0$ or $\Omega_k = 0$, as shown in Example 5.3 and in Problem (5.11), and illustrated in Fig. 5.6. The result (5.15) obviously follows

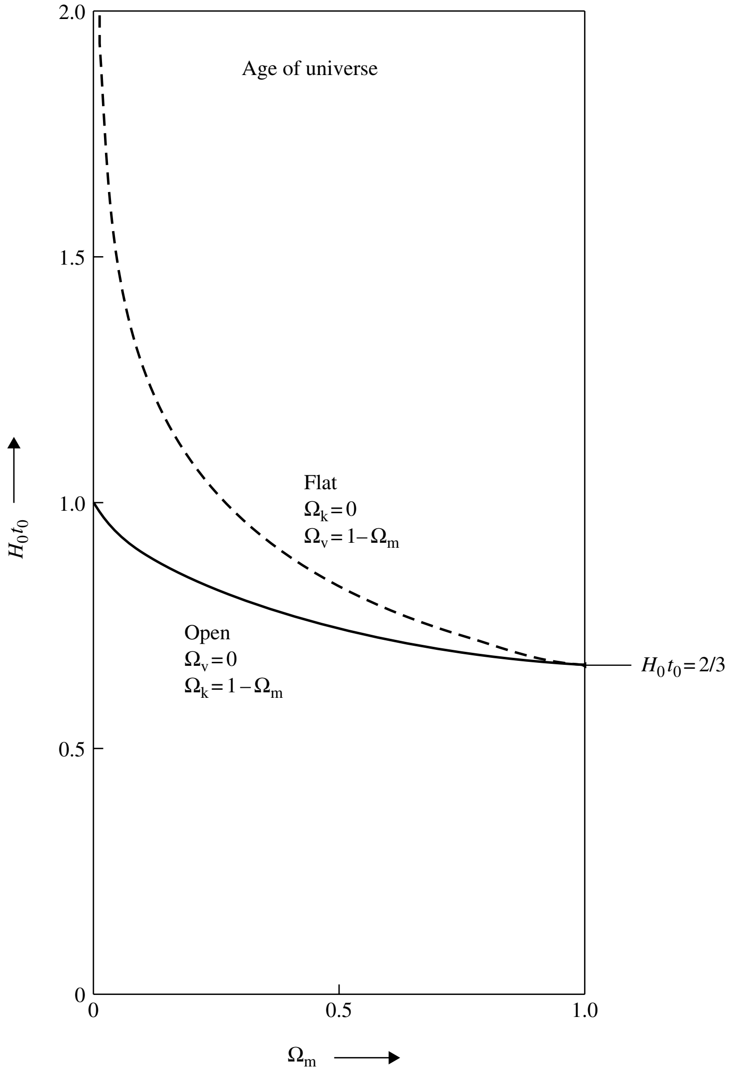


Fig. 5.6 Plot of the age of the universe versus the parameter Ω_m, the ratio of the matter density to the critical density. The solid curve is for an open universe, in which the curvature term $\Omega_k = 1 - \Omega_m$, and radiation and vacuum energy terms are assumed to be zero. The dashed curve is for a flat universe ($\Omega_k = 0$), in which radiation energy is neglected and the vacuum energy $\Omega_v = 1 - \Omega_m$. The present best estimate relates to the flat universe with $\Omega_m = 0.24$. The curves have been calculated from the analytical expressions in Example 5.3 and in Problem 5.11.

when radiation, vacuum, and curvature terms are all zero, and the universe is flat and matter dominated.

Example 5.3 *Estimate the age of a flat universe ($k = 0$) if radiation is neglected and it is presently made up of matter with Ω_m= 0.24 and vacuum energy with* $\Omega_\Lambda = 0.76$.

In this case, the above integral (5.39) becomes

$$H_0 t_0 = \int_0^\infty \frac{\mathrm{d}z}{(1+z)\left[\Omega(1+z)^3 + (1-\Omega)\right]^{1/2}}$$

where $\Omega \equiv \Omega_m(0)$ and $\Omega_\Lambda(0) = (1 - \Omega)$. The integral is readily evaluated with the substitution $\Omega(1+z)^3/(1-\Omega) = \tan^2\theta$, when it transforms to the integral $\int \mathrm{d}\theta/\sin\theta = \ln[\tan(\theta/2)]$. Finally one obtains

$$H_0 t_0 = \left[\frac{1}{(3A)}\right] \ln\left[\frac{(1+A)}{(1-A)}\right]$$

where $A = (1-\Omega)^{1/2}$. For $\Omega = 0.24$, $(1-\Omega) = 0.76$, one finds $H_0 t_0 = 1.026$, so that $t_0 = 1.026/H_0 = 13.95 \pm 0.4$ Gyr. The vacuum term has thus increased the age over the value (5.15).

The *radius of the observable universe* is determined by the distance to the optical horizon, beyond which no light signals could reach the Earth at the present time. As time evolves, this distance increases as more parts come inside the horizon. In a static, flat universe, the horizon distance would simply be the product

$$D_\mathrm{H} = ct_0 = 4.2 \text{ Gpc} \tag{5.40}$$

where t_0 is the age described above. Clearly, a somewhat larger value would be obtained in an expanding universe. In the FLRW model of an isotropic and expanding universe with uniform curvature, introduced in Section 2.9, the true coordinate distance to any point at time t is given by $D(t) = rR(t)$ as in (5.8), where r is the co-moving coordinate distance (i.e. the distance measured on a scale expanding with the Hubble expansion) and $R(t)$ is the universal scale factor. Of course, neither of these quantities can be measured directly. We need to express them in terms of measurable quantities, namely the Hubble parameter and the redshift z.

From (2.31) the line element in the FLRW model is

$$\mathrm{d}s^2 = c^2\mathrm{d}t^2 - R(t)^2\left[\frac{\mathrm{d}r^2}{\left(1-kr^2\right)} + r^2\mathrm{d}\theta^2 + r^2\sin^2\theta\,\mathrm{d}\varphi^2\right] \tag{5.41}$$

Consider the path of a photon to or from some distant object at fixed (θ, φ), for which we know from Section 2.2 that $\mathrm{d}s^2 = 0$. With $R(t) = R(0)/(1+z)$ we find from (5.41)

$$c(1+z)\mathrm{d}t = \frac{R(0)\mathrm{d}r}{\sqrt{(1-kr^2)}}$$

and from (5.38)

$$c(1+z)\mathrm{d}t = -\frac{c\mathrm{d}z}{H}$$

Hence

$$R(0)\int_0^r \frac{\mathrm{d}r}{\sqrt{1-kr^2}} = -\int \frac{c\mathrm{d}z}{H} = \frac{cI(z)}{H_0} \tag{5.42}$$

where from (5.37)

$$I(z) = \int_0^z \frac{\mathrm{d}z}{\left[\Omega_m(0)(1+z)^3 + \Omega_r(1+z)^4 + \Omega_\Lambda(0) + \Omega_k(0)(1+z)^2\right]^{1/2}} \tag{5.43}$$

Carrying out the integration over r on the left-hand side of (5.42), one obtains for the three possible values of k:

$$\begin{aligned} cI(z)/H_0 &= R(0)\sin^{-1} r && k=+1 && \text{closed} \\ &= R(0)\sinh^{-1} r && k=-1 && \text{open} \\ &= R(0)r && k=0 && \text{flat} \end{aligned} \tag{5.44a}$$

So the present true coordinate distance of our object at redshift z is

$$\begin{aligned} D(z) = rR(0) &= \left[\frac{c}{(H_0 Q)}\right]\sin\left[I(z)Q\right] && k=+1 && \text{closed} \\ &= \left[\frac{c}{(H_0 Q)}\right]\sinh\left[I(z)Q\right] && k=-1 && \text{open} \\ &= \left[\frac{cI(z)}{H_0}\right] && k=0 && \text{flat} \end{aligned} \tag{5.44b}$$

where $Q = |\Omega_k(0)|^{1/2}$. The horizon distance D_{H} is then obtained by setting the upper limit of integration in (5.43) as $z = \infty$. As an example, for a flat, matter-dominated universe, that is, $\Omega_m(0) = 1$ and all other contributions set to zero, one obtains $D_{\mathrm{H}} = 2c/H_0$, while for the case of a flat radiation-dominated universe with $\Omega_r(0) = 1$, $D_{\mathrm{H}} = c/H_0$.

For the values of the contributions to the closure parameter $\Omega_{\mathrm{tot}} = 1$ quoted above, that is, $\Omega_m(0) = 0.24$, $\Omega_\Lambda(0) = 0.76$, $\Omega_r(0) = \Omega_k(0) = 0$, the integral (5.43) has to be evaluated numerically, with the result that the horizon distance or visible radius of the universe become

$$D_{\mathrm{H}} \sim 3.3\frac{c}{H_0} \sim 14 \;\; \mathrm{Gpc} \tag{5.45}$$

Of course, if the dark energy term, here identified with vacuum energy, is z-dependent, this result would change.

5.7 The deceleration parameter: the effects of vacuum energy/cosmological constant

One can express the time dependence of the expansion parameter as a Taylor series:

$$R(t) = R(0) + \dot{R}(0)\,(t - t_0) + \left(\frac{1}{2}\right)\ddot{R}(0)\,(t-t_0)^2 + \cdots$$

or

$$\frac{R(t)}{R(0)} = 1 + H_0\,(t - t_0) - \left(\frac{1}{2}\right) q_0 H_0^2\,(t-t_0)^2 + \cdots$$

where the *deceleration parameter*, which can be time dependent, is defined as

$$\begin{aligned} q &= -\ddot{R}R/\dot{R}^2 \\ &= \left[\frac{4\pi \mathrm{G}}{\left(3c^2H^2\right)}\right]\left[\rho c^2 + 3P\right] \end{aligned} \tag{5.46a}$$

from (5.20). Inserting the values of ρ and P for the components in Table 5.2, it is straightforward to show that this dimensionless parameter has the value

$$q = \frac{\Omega_m}{2} + \Omega_r - \Omega_\Lambda \tag{5.46b}$$

Today $\Omega_m \gg \Omega_r$, so that if Ω_Λ could be neglected, a flat universe would have $\Omega = \Omega_m = 1$ and $q = 0.5$, that is, the universal expansion must be decelerating because of the retarding effects of the gravitational attraction of matter. In fact, early attempts to measure q seemed to give results consistent with this value (within large errors). We may note that if Ω_Λ is large enough however, $q < 0$ and the expansion would be *accelerating*, the vacuum energy having the same effect as a gravitational repulsion. As mentioned above, surveys on Type 1A supernovae at high redshifts, treating them as 'standard candles', appear to indicate that q is indeed negative, as described in Chapter 7. These surveys show that several billion years ago, that is, for redshifts $z > 1$, the universe *was* indeed decelerating, but that more recently this deceleration has been replaced by an acceleration. We note here from (5.46) that an *empty universe*, that is one with $\Omega_m = \Omega_\Lambda = \Omega_r = 0$, and hence $\Omega_k = 1$, is neither accelerating nor decelerating, with H independent of time (see also (5.29)). Thus an empty universe is the yardstick against which in Chapter 7 we judge that a particular model results in acceleration or retardation.

5.8 CMB radiation

One of the major discoveries in astrophysics was made in 1965 by Penzias and Wilson. While searching for cosmic sources of radio waves at approximately 7 cm wavelength, they discovered an isotropic background of microwave radiation. Although they were unaware of it, this had been predicted by Gamow many years before, as a relic of the Big Bang, a photon fireball cooled by expansion to a temperature of a few degrees kelvin. Figure 5.7 shows data on the spectral distribution of radiation recorded originally by the Cosmic Background Explorer (COBE) satellite (Smoot *et al.* 1990). Satellite and balloon-borne detectors as well as ground-level interferometers have since then mapped out the spectrum over an enormous range of wavelengths, from 0.05 to 75 cm. Recent data show very precise agreement with the spectrum expected from a black body at a temperature of 2.725 ± 0.001 K; indeed, the cosmic microwave spectrum is *the* black body spectrum *par excellence*. It proves among other things that at the time the radiation last interacted significantly with matter, it was in thermal equilibrium with it. In fact, the CMB observed today originated when matter and radiation decoupled as the universe expanded and cooled, some 380,000 years after the Big Bang.

Assuming that matter has been conserved, the matter density of the universe can be expected to vary as $\rho_m \propto R^{-3}$. On the other hand, the density of radiation, assuming it to be in thermal equilibrium, will vary with temperature as $\rho_r \propto T^4$ (Stefan's Law). Since there is no absolute scale of distance, the wavelength of the radiation on the truly cosmic scale associated with the Hubble expansion can only be proportional to the expansion factor R. Thus the frequency $\nu = c/\lambda$ and the mean energy per photon will both be proportional

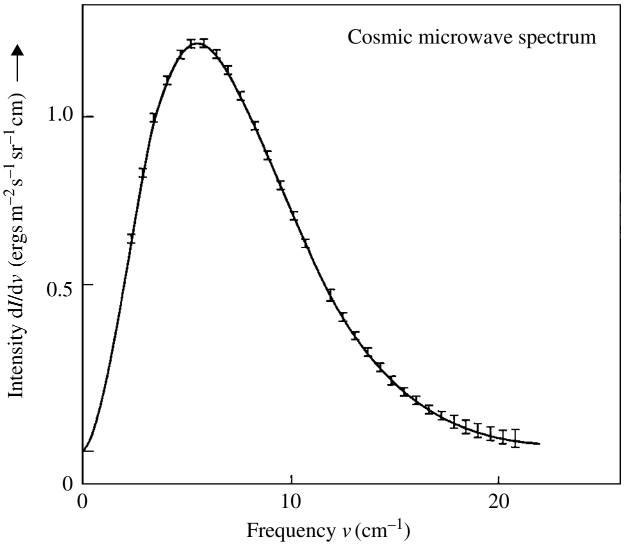


Fig. 5.7 Data on the spectral distribution of the cosmic microwave radiation obtained from the COBE satellite experiment. The experimental points show the results of the early experiments in 1990. When recent satellite data and those from balloon -borne experiments are combined, a very exact fit to a black body spectrum is obtained with $T = 2.725 \pm 0.001$ K and $kT = 0.235$ meV (milli-electron volts) as shown by the curve (Fixen *et al.* 1996). The present experimental errors are actually less than the thickness of this curve.

to R^{-1}. While the number of photons varies as $1/R^3$, the energy density of the radiation will vary as $1/R^4$, as indicated in Table 5.2. The extra factor of $1/R$ in the energy density, as compared with non-relativistic matter, simply arises from the redshift, which in fact will apply to any relativistic particles and not just to photons, provided of course that those particles are distributed uniformly on the same cosmological scale as the microwave photons. At the early times we are discussing here, the vacuum energy, which is assumed to be independent of R, would have been totally negligible and we can just forget it.

Thus, while the matter density of the universe dominates over radiation today, in the olden days and at low values of R, radiation must have been dominant. In that case, the second term on the right-hand side of (5.11) can be neglected in comparison with the first, varying as $1/R^4$. Then

$$\dot{R}^2 = \left(\frac{8\pi G}{3}\right)\rho_r R^2$$

Furthermore, since $\rho_r \propto R^{-4}$,

$$\frac{\dot{\rho}_r}{\rho_r} = -\frac{4\dot{R}}{R} = -4\left(\frac{8\pi G\rho_r}{3}\right)^{1/2}$$

which on integration gives for the energy density

$$\rho_r c^2 = \left(\frac{3c^2/32\pi G}{t^2}\right) \tag{5.47}$$

For a photon gas in thermal equilibrium

$$\rho_r c^2 = \frac{4\sigma T^4}{c} = \pi^4 (kT)^4 \left(\frac{g_\gamma/2}{15\pi^2\hbar^3c^3}\right) \tag{5.48}$$

where k is here the Boltzmann constant. (This should not to be confused with the curvature parameter, also denoted by k; since the Boltzmann constant will always occur multiplied by the temperature T.) σ is the Stefan–Boltzmann constant and $g_\gamma = 2$ is the number of spin substates of the photon. From these last two equations we obtain a relation between the temperature of the radiation and the time of expansion:

$$kT = \frac{\left[\left(45\hbar^3c^5 \Big/ 16\pi^3 \mathrm{G} g_\gamma\right)^{1/4}\right]}{t^{1/2}} = 1.307 \ \frac{\text{MeV}}{t^{1/2}} \tag{5.49}$$

where t is in seconds. The corresponding value of the temperature itself is

$$T = 1.52 \times 10^{10} \frac{K}{t^{1/2}}.$$

Since T falls as $1/R, R$ increases as $t^{1/2}$ while the temperature falls as $1/t^{1/2}$. Hence, the universe started out as a hot Big Bang.

From (5.49) we may roughly estimate the energy of the radiation today, that is for $t_0 \sim 14$ Gyr $\sim 10^{18}$ s. It is $kT \sim 1$ meV (milli-electron volt), corresponding to a temperature of a few degrees on the Kelvin scale. This will in fact be an overestimate since the radiation has cooled more quickly, as $1/t^{2/3}$, during the later, matter-dominated era (see Fig. 5.10).

Observation of microwave molecular absorption bands in distant gas clouds has made it possible to estimate the temperature of the background radiation at earlier times, when the wavelength would have been reduced, and the temperature increased, by the redshift factor $(1+z)$. This dependence on redshift has been experimentally verified up to values of $z \approx 3$.

Let us now compare the observed and expected energy densities of radiation. The spectrum of black body photons of energy $E = pc = h\nu$ is given by the Bose–Einstein (BE) distribution, describing the number of photons per unit volume in the momentum interval $p \to p + \mathrm{d}p$. Including $g_\gamma = 2$ as the number of spin substates of the photon, this is

$$N(p)\mathrm{d}p = \frac{p^2 \mathrm{d}p}{\pi^2\hbar^3 \left\{\exp\left(E/kT\right) - 1\right\}} \ \left(\frac{g_\gamma}{2}\right) \tag{5.50}$$

In discussing the BE distribution, and later, the Fermi–Dirac (FD) distribution, it will be useful to note the following integrals, from $x = 0$ to $x = \infty$:

$$\begin{aligned} \text{BE}: \quad & \int \frac{x^3 \mathrm{d}x}{(e^x - 1)} = \frac{\pi^4}{15}; \quad \int \frac{x^2 \mathrm{d}x}{(e^x - 1)} = 2.404 \\ \text{FD}: \quad & \int \frac{x^3 \mathrm{d}x}{(e^x + 1)} = \frac{7}{8} \times \frac{\pi^4}{15}; \quad \int \frac{x^2 \mathrm{d}x}{(e^x + 1)} = \frac{3}{4} \times 2.404 \end{aligned} \tag{5.51}$$

The total energy density integrated over the spectrum is then readily calculated to have the value ρ_r in (5.48). The number of photons per unit volume is

$$N_\gamma = \left(\frac{2.404}{\pi^2}\right)\left(\frac{kT}{\hbar c}\right)^3 = 411\left(\frac{T}{2.725}\right)^3 = 411 \ \text{cm}^{-3} \tag{5.52}$$

while the energy density from (5.48) is

$$\rho_r c^2 = 0.261 \;\; \mathrm{MeV\,m^{-3}} \tag{5.53}$$

the equivalent mass density being

$$\rho_r = 4.65 \times 10^{-31} \;\; \mathrm{kg\,m^{-3}}$$

and from (5.26)

$$\Omega_r(0) = 4.84 \times 10^{-5} \tag{5.54}$$

some four orders of magnitude less than the present estimated matter density in (5.33).

5.9 Anisotropies in the microwave radiation

The temperature of the microwave radiation shows a small anisotropy, of order 10^{-3}, attributed to the 'peculiar velocity' $v = 370$ km s^{-1} of the Solar System (towards the Virgo cluster) with respect to the (isotropic) radiation. It is given by the Doppler formula (2.36):

$$T(\theta) = T(0)\left[1 + \left(\frac{v}{c}\right)\cos\theta\right] \tag{5.55}$$

where θ is the direction of observation with respect to the velocity v. Figure 5.8 shows (magnified in contrast by 400 times) the 'hot' ($\theta = 0$) and 'cold' ($\theta = \pi$) features of the dipole, as well as the (infrared) emission from the galaxy, showing as a broad central band. After the dipole contribution and the galactic emission are removed, a polynomial analysis of the distribution shows that there are quadrupole ($l = 2$) and higher terms, up to at least $l = 1000$, involving tiny but highly significant anisotropies at the 10^{-5} level. These turn out to be of fundamental importance, reflecting fluctuations in density and temperature in the early universe which seeded the large-scale structures observed today. These matters are discussed in detail in Sections 8.13 to 8.16.

As indicated in Section 5.12, the microwave radiation, previously in equilibrium with atomic and ionized hydrogen, decoupled from baryonic matter at $z \sim 1100$, when the universe was about 400,000 years old. That would have

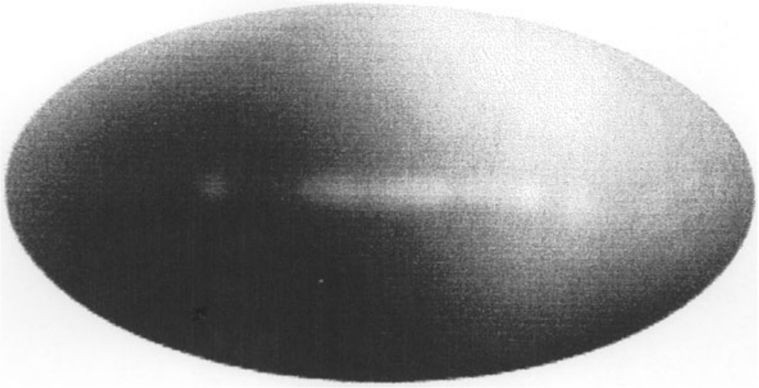


Fig. 5.8 Plot of the angular distribution of the microwave background radiation, showing the dipole dependence of (5.55) due to the velocity of the Earth relative to the isotropic radiation, plus the infrared emission from the Milky Way, showing as a broad central band. The angular dependence shown has been enhanced some 400 times from the actual value, of order 10^{-3}.

been the epoch of 'last scattering', if the interstellar gas (mostly hydrogen and helium) remained unionized. However, it appears that when z fell below about 12 (the end of the so-called dark ages), the first stars had formed and commenced re-ionization of the intergalactic medium, by the ultraviolet radiation they emitted. Thus the microwave radiation, on its passage through the interstellar medium to the observer, would then undergo Thomson scattering from electrons in the plasma. It is, however, a small effect (see Section 8.14 et seq).

Example 5.4 *Calculate the mean quantum energy and the corresponding wavelength of the cosmic microwave photons for a temperature of $T = 2.725\,K$.*

The original discovery of cosmic microwave radiation was made with receivers tuned to 7.3 cm wavelength. What fraction of the photons would have wavelengths in excess of 7.3 cm?

From (5.50) and (5.51) the mean photon energy is $\pi^4 kT/(15 \times 2.404) = 2.701\ kT = 6.34 \times 10^{-4}$ eV. The corresponding wavelength is $\lambda = hc/hv = 0.195$ cm.

At large wavelengths the curly bracket in (5.50) can be approximated by E/kT if $E/kT \ll 1$. The fraction of photons with quantum energies below $\varepsilon = E/kT$ is then easily shown to be $F = \left(\varepsilon/kT\right)^2/(2 \times 2.404)$, which for wavelengths above 7.3 cm is equal to 1.06×10^{-3}.

5.10 Particles and radiations in the early universe

The relation (5.49) for the temperature of the early universe as a function of time applies for radiation consisting of photons (with $g_\gamma = 2$). Relativistic fermions, that is, quarks and leptons, assuming that they are stable enough, would also contribute to the energy density. For a fermion gas, the FD distribution for the number density analogous to (5.50) is

$$N(p)\mathrm{d}p = \frac{p^2 \mathrm{d}p}{\pi^2 \hbar^3 \left\{\exp\left(E/kT\right) + 1\right\}} \left(\frac{g_f}{2}\right) \tag{5.56}$$

where $E^2 = p^2c^2 + m^2c^4$, m is the fermion mass and g_f is the number of spin substates. In the relativistic limit, $kT \gg mc^2$ and $E = pc$, the total energy density, in comparison with (5.48), is given by (see (5.51)):

$$\rho_f c^2 = \left(\frac{7}{8}\right) \pi^4 (kT)^4 \frac{(g_f/2)}{15\pi^2 \hbar^3 c^3} \tag{5.57}$$

Thus, for a mixture of extreme relativistic bosons b and fermions f, the energy density in (5.48) is found by replacing g_γ by a factor g^* where

$$g^* = \sum g_b + \left(\frac{7}{8}\right) \sum g_f \tag{5.58}$$

and the summation is over all types of relativistic particles and antiparticles which contribute to the energy density of radiation in the early universe.

Of course, at very early times when the temperature was high enough for their creation, all types of elementary quarks, leptons, and bosons, plus their antiparticles, would have been present in the primordial 'soup' in which the various components would have been in thermal equilibrium. On the basis of the fundamental particles we know today, the number of degrees of freedom (charge, spin, and colour substates) of the fermions would be 90, and that of the gauge bosons 28.

To understand these rather big numbers, recall that the bosons include the massless photon of spin 1, occurring in 2 spin states since, according to relativistic invariance, a massless particle of spin J can have only two substates, $J_z = \pm J$; the massless gluon also of spin 1, 2 spin substates, and 8 substates of colour; the massive bosons W^+, W^-, and Z^0, again of spin 1 but since they are massive, contribute $2J + 1 = 3$ spin substates each; and finally the Higgs scalar spin 0 boson of the electroweak theory described in Chapter 3, bringing the total to 28. The fermions include the quarks, occurring in 6 flavour states, 3 colour, and 2 spin substates, plus their antiparticles, totalling 72 states altogether; the charged leptons in 3 flavour and 2 spin substates, plus their antiparticles, that is a total of 12 states; and finally the neutral leptons (neutrinos) in 3 flavours but only one spin substate each. Including antiparticles the neutrinos contribute 6 degrees of freedom, making 90 fermion and antifermion states in total. Of course, in this tally we have counted only the known fundamental particles. If supersymmetry is valid, for example, the number of states will be approximately doubled. We note here that, for values of kT very much larger than any of the particle masses, then inserting $g_b = 28$ and $g_f = 90$, the value of $g^* = 106.75$, as shown in Fig. 5.9.

As the expansion proceeded and the temperature fell, the most massive particles, such as the top quark and the W and Z bosons would have been rapidly lost by decay (in less than 10^{-23} s) and not replenished once $kT \ll Mc^2$ where M is the particle mass. After kT fell below the strong quantum chromodynamics (QCD) scale parameter ~ 200 MeV, the remaining quarks, antiquarks, and gluons would no longer exist as separate components of a plasma but as quark bound states, forming the lighter hadrons such as pions and nucleons. However, all hadrons except protons and neutrons would be too short-lived to exist beyond the first few nanoseconds. Similarly, the charged muon and tauon leptons would decay within the first microsecond or so. Once kT had fallen below about 20 MeV, that is after the first few milliseconds, most of the nucleons and antinucleons would also have annihilated to radiation, as discussed in Chapter 6. The surviving number of nucleons in fact amounts to only about one billionth of the number of photons. This would leave, apart from the photons, the electrons e^- and the ν_e, ν_μ, and ν_τ neutrinos, plus their antiparticles, giving in (5.58) $\Sigma g_f = 4 + 2 + 2 + 2$ (recalling two spin states each for electrons and positrons, but only one for the neutrinos or antineutrinos). With $g_b = 2$ for the photon this results in a value $g^* = 43/4$. The effect is to multiply the expression for kT on the extreme right-hand side of (5.49) by a factor $\left(g^*/2\right)^{-1/4}$, which in this case has the value 0.66. Note that this result applies for values of kT between about 20 MeV and 5 MeV, as is shown in Fig. 5.9.

From the formulae of the last two sections we may also express the Hubble parameter $H(t)$ in terms of the temperature T in the radiation-dominated era of

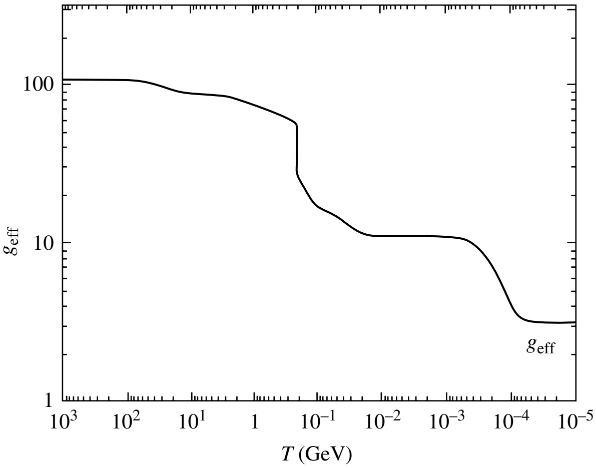


Fig. 5.9 Plot of the quantity g^* in (5.58)—here termed g_{eff}—measuring the number of degrees of freedom, against the temperature kT (after Kolb and Turner 1990).

the early universe. Since in this era, $\rho \propto R^{-4}$, it follows from (5.47) that

$$H(t) = \frac{\dot{R}}{R} = -\frac{\dot{\rho}}{4\rho} = \frac{1}{(2t)}$$

and from (5.49)

$$\begin{aligned} H(t) &= \left[\frac{4g^*\pi^3 \mathrm{G}}{45\hbar^3 c^5}\right]^{1/2} (kT)^2 \\ &= \frac{\left(4\pi^3\, g^*/45\right)^{1/2}}{M_{\text{PL}}\hbar c^2} \times (kT)^2 \\ &= 1.66 g^{*^{1/2}} \frac{(kT)^2}{M_{\text{PL}}\hbar c^2} \end{aligned} \tag{5.59}$$

where in the second line the Newtonian constant is expressed in terms of the Planck mass, that is $G = \hbar c/M_{\text{PL}}^2$ (see Table 1.5).

Example 5.5 *Estimate the time required for the universe to increase its size by 10% during the radiation era, for values of kT = 100 MeV and* $g^* = 20$.

Since $H = (1/R)\,\mathrm{d}R/\mathrm{d}t$, the time required (assuming that H is constant over a short period) is found on integration to be $t = (\ln 1.1)/H$. From (5.59), with $M_{\text{PL}} = 1.22 \times 10^{19}$ GeV/c^2 and kT expressed in GeV:

$$H(t) = 2.07 \times 10^5 g^{*^{1/2}} (kT)^2 \ \ \mathrm{s}^{-1}$$

Substituting for kT and g^* we find $H = 9.25 \times 10^3$ s^{-1} and $t = 10.3$ μs.

5.11 Photon and neutrino densities

When kT is down to a few MeV, the only surviving relativistic particles (radiation) would consist of neutrinos ν_e, ν_μ, and ν_τ and their antiparticles, together with electrons, positrons, and photons. These would have been by far the most prolific particles in the early universe, since as shown in Chapter 6, the number of the only other stable particles, neutrons and protons, would have been comparatively small. The light leptons and photons would have been present in comparable numbers, according to the equilibrium reactions

$$\gamma \leftrightarrow \mathrm{e}^+ + \mathrm{e}^- \leftrightarrow \nu_i + \bar{\nu}_i \tag{5.60}$$

where $i = e, \mu, \tau$. As indicated in Section 3.6, the cross-section for electron–positron annihilation to a neutrino–antineutrino pair is a weak process with a cross-section of order $\sigma \sim G_\mathrm{F}^2 s/6\pi$, where s is the square of the CMS energy (see Example 3.2).

The collision rate for this reaction is $W = \langle \rho\sigma v\rangle$ where ρ is the number density of electrons or positrons, and v is their relative velocity. Since $\rho \sim T^3$ is the number density of relativistic particles at temperature T, the rate $W \sim sT^3 \sim T^5$, while the universal expansion rate in the radiation-dominated universe is $H \sim T^2$ as in (5.59). Hence as T falls during the expansion, neutrinos must start to decouple as soon as $W < 1/H$. Inserting numerical values (see Problem 5.12) one finds for the critical temperature $kT \sim 3$ MeV. So for $t > 1$ s, neutrinos are therefore decoupled, and the neutrino fireball expands and cools independently of the other particles or radiation (apart, of course, from the universal redshift).

The number density of the neutrinos in (5.60) will be comparable with that of the photons. However, when $kT < 1$ MeV, the photons will be boosted by the annihilation process $\mathrm{e}^+ + \mathrm{e}^- \rightarrow \gamma + \gamma$, which converts the energy content of electrons and positrons into photons. The entropy per unit volume of the particle gas will be $S = \int \mathrm{d}Q/T$ where Q is the energy content per unit volume of photons, electrons, and positrons at temperature T, and from (5.48) and the integrals (5.51)

$$S = \left(\int \frac{4aT^3\mathrm{d}T}{T}\right) \times \left(1 + \frac{7}{8} + \frac{7}{8}\right) = \left[\frac{4aT^3}{3}\right] \times \frac{11}{4} \tag{5.61}$$

where $a = 4\sigma_\mathrm{st}/c$ is the radiation constant and σ_st is Stefan's constant. After annihilation the photons will have attained a temperature T_1 with entropy

$$S_1 = \left(\frac{4a}{3}\right) T_1^3 \tag{5.62}$$

but since the expansion is adiabatic (isentropic) $S_1 = S$ so that

$$T_1 = \left(\frac{11}{4}\right)^{1/3} T \tag{5.63}$$

So if the temperature of the microwave photons was T_γ, that of the relic neutrinos, which received no boost would be

$$T_\nu = \left(\frac{4}{11}\right)^{1/3} T_\gamma \tag{5.64}$$

Consequently, the value of g^* used to multiply the factor $(kT)^4$ in (5.48) would have the value

$$g^* = g_\gamma + \left(\frac{7}{8}\right) g_\nu \left(\frac{T_\nu}{T_\gamma}\right)^4 \tag{5.65}$$

and with $g_\nu = 6$ and $g_\gamma = 2$ we obtain

$$g^* = 2 + \left(\frac{21}{4}\right) \times \left(\frac{4}{11}\right)^{4/3} = 3.36 \tag{5.66}$$

applying to the region $kT \ll 1$ MeV, and shown in Fig. 5.9. The relic neutrinos and photons do not interact further, they suffer exactly the same redshift as the universe expands and cools, and so the relative numbers today will be the same as indicated above. From the number of microwave photons observed today in (5.52), it is easy to show that the number of microwave neutrinos plus antineutrinos should be

$$N_\nu = \left(\frac{3}{11}\right) N_\gamma = 113 \text{ cm}^{-3} \tag{5.67}$$

per neutrino flavour, to be compared with a number density of 411 cm^{-3} for the microwave photons. From (5.63), the temperature $T_\nu = 1.95$ K, compared with the value $T_\gamma = 2.73$ K for the photons. This result, however, assumes that the neutrinos are, even today, extreme relativistic. With $kT_\nu \sim 0.17$ meV, this cannot be true for all (or perhaps any) neutrino flavours, as the mass differences in (4.12) of 10–50 meV/c^2 indicate. The role of relic neutrinos is further discussed in Chapter 7.

5.12 Radiation and matter eras: the decoupling of matter and radiation

From the above formulae, for example (5.49), it is apparent that at early times in the universe when the temperature and particle densities were extremely high, the various types of elementary fermions and bosons would have been in thermal equilibrium and present in comparable numbers, provided $kT \gg Mc^2$, so that even the most massive particles could have been created. The condition for thermal equilibrium to apply is that the time between collisions should be much shorter than the age t of the universe. Otherwise, there is just not enough time to have had enough collisions to set up equilibrium ratios. The collision rate of a particle will be $W = \langle N v \sigma \rangle$ where N is the density of other particles with which it collides, σ is the cross-section per collision, and an average is taken over the distribution in relative velocity v. So one requires that $W \gg t^{-1}$.

Eventually, particles may fall out of equilibrium, as the universe expands and the temperature decreases. For example, the cross-section may depend on energy and become so small at low temperature that W falls below t^{-1}, and those particles therefore decouple from the rest. We say that they are ‘frozen-out’. As indicated in the previous section, this is the case for the weak reaction

$$e^+ + \mathrm{e}^- \leftrightarrow \nu + \bar{\nu}$$

for $kT \ll 3$ MeV, that is when $t > 1$ s. So after that time, the neutrino fireball is decoupled from matter and expands independently. Although these primordial neutrinos have essentially no further electroweak interactions with matter, they of course have gravitational interactions and play an important role in the clustering of matter, which will ultimately result in the large-scale structure of galaxies, superclusters, and so forth as will be discussed in Chapter 8.

Particles may also decouple if they are massive, even if the production cross-section is large. For example, this will happen for the reversible process

$$\gamma + \gamma \leftrightarrow p + \bar{p}$$

when $kT \ll M_p c^2$. This reaction will be discussed in the next chapter.

For some 10^5 years after the Big Bang, baryonic matter, consisting largely of protons, electrons, and hydrogen atoms, was in equilibrium with the photons, via the reversible reaction

$$\mathrm{e}^- + p \leftrightarrow H + \gamma \tag{5.68}$$

where, in the forward process, a hydrogen atom is formed, in the ground state or in an excited state, and in the reverse process, a hydrogen atom is ionized by the radiation, the resultant protons and electrons forming a plasma. At thermal equilibrium, the ratio of ionized to unionized hydrogen is a constant depending on the temperature T. We are interested in what happens as the temperature falls and $kT < I$, the ionization potential of hydrogen ($I = 13.6\,\mathrm{eV}$). Clearly, the rate for the forward reaction is proportional to the product of the densities N_e and N_p of the electrons and protons, while the back reaction rate will be proportional to the number N_H of hydrogen atoms per unit volume. (The number of photons is enormous by comparison, so their number is unaffected by the reaction.) Hence

$$\frac{N_\mathrm{e} N_\mathrm{p}}{N_\mathrm{H}} = f(T) \tag{5.69}$$

The number of bound states available to an electron will be $g_\mathrm{e}\, g_\mathrm{n}$ where $g_\mathrm{e} = 2$ is the number of spin substates and $g_\mathrm{n} = n^2$ is the number of bound states in a hydrogen atom with principal quantum number n and energy E_n. The probability that an electron is bound in a state of energy E_n is found by multiplying by the Boltzmann factor, so that it is $g_\mathrm{e} g_\mathrm{n} \exp(-E_\mathrm{n}/kT)$. Summed over ground ($n = 1$) and excited states ($n > 1$) of the H atom, the probability to find an electron in a *bound* state is therefore

$$P_\mathrm{bound} = g_\mathrm{e} \sum g_\mathrm{n} \exp\left(\frac{-E_n}{kT}\right)$$

If we write $-E_n = -E_1 - (E_n - E_1)$ where $-E_1 = I$, the ionization potential, then

$$P_\mathrm{bound} = g_\mathrm{e} Q \exp\left(\frac{I}{kT}\right) \tag{5.70}$$

where

$$Q = \sum n^2 \exp\left[-\frac{(E_n - E_1)}{kT}\right]$$

Since $(E_n - E_1)/kT \gg 1$ for all values of $n > 1$, the excited states make little contribution and $Q \approx 1$.

The probability that our electron is in an *unbound* state of kinetic energy $E \to E + \mathrm{d}E$ is

$$P_{\text{unbound}} = g_{\text{e}} \left(\frac{4\pi p^2\, dp}{h^3} \right) \exp\left(\frac{-E}{kT} \right)$$

where $4\pi p^2 \mathrm{d}p/h^3$ is the number of quantum states per unit volume in the interval $p \to p + \mathrm{d}p$ and $\exp(-E/kT)$ is the probability that any such state will be occupied by an electron of kinetic energy $E = p^2/2m$, where m is the electron mass. Here we have assumed the electron is non-relativistic and that $E \gg kT$, so that the FD occupation probability in (5.56) reverts to the classical Boltzmann factor $\exp(-E/kT)$. The probability that the electron will be unbound with *any* energy $E > 0$ is found by integrating over E, with the result

$$P_{\text{unbound}} = g_{\text{e}} \left(\frac{2\pi mkT}{h^2} \right)^{3/2} \tag{5.71}$$

(see Problem 5.3). Comparing the relative probabilities in (5.70) and (5.71), and using (5.69), the ratio of unbound (ionized) to bound (unionized) states is

$$\frac{N_{\text{p}}}{N_{\text{H}}} = \frac{N_{\text{H}^+}}{N_{\text{H}}} = \left(\frac{1}{N_{\text{e}}} \right) \left(\frac{2\pi mkT}{h^2} \right)^{3/2} \exp\left(\frac{-I}{kT} \right) \tag{5.72}$$

The total number of baryons per unit volume is $N_{\text{B}} = N_{\text{p}} + N_{\text{H}}$, so that if x represents the fraction of hydrogen atoms which are ionized, then $N_{\text{e}} = N_{\text{p}} = xN_{\text{B}}$ and $N_{\text{H}} = (1 - x)N_{\text{B}}$, hence

$$\frac{x^2}{(1-x)} = \left(\frac{1}{N_{\text{B}}} \right) \left(\frac{2\pi mkT}{h^2} \right)^{3/2} \exp\left(\frac{-I}{kT} \right) \tag{5.73}$$

called the *Saha equation*. Inserting some typical numbers, the reader can easily demonstrate from this formula that for kT between 0.35 eV (4000 K) and 0.25 eV (3000 K), x drops catastrophically, and so radiation and matter must *decouple* at around this temperature. (One can also phrase this epoch as the time of the *recombination* of electrons and protons, to form hydrogen atoms.) A value of $kT = 0.30$ eV is in fact a good guess for the decoupling temperature. Comparing this with the value $kT_0 = 2.35 \times 10^{-4}$ eV ($T_0 = 2.73$ K) of the microwave radiation at the present day, the value of the redshift at the time of decoupling will have the value

$$(1 + z)_{\text{dec}} = \frac{R(0)}{R_{\text{dec}}} = \frac{kT_{\text{dec}}}{\mathrm{kT}_0} \approx 1250 \tag{5.74}$$

Corrections to this result are needed, since a photon emitted upon recombination of one atom can almost immediately ionize another atom, and the level of ionization could therefore be underestimated. In fact it turns out that slower, two-step processes involving more than one photon are of importance and the value of the redshift deduced from these more detailed calculations is found to be

$$(1 + z)_{\text{dec}} = 1100 \tag{5.75}$$

Of course, the decoupling does not all occur at a single value of the redshift. The distribution of ‘last scattering’events is spread out over an r.m.s. variation of

$\Delta z \sim 80$. In this treatment, we have assumed thermal equilibrium throughout, and indeed it can be shown that the above reactions do stay in equilibrium and are not totally 'frozen-out' by the Hubble expansion until z has fallen below z_{dec}. The ionization fraction surviving at the eventual freeze-out is, however, very small, $x < 10^{-3}$.

After decoupling, matter becomes transparent to the CMB radiation, and the formation of atoms and molecules can begin in earnest. Equally important, some vital properties of this radiation, including the very small but very important spatial variations in temperature which can be observed today, are therefore very close to what they were at the 'epoch of the last scattering'. These measurements of temperature fluctuations in very small angular ranges (of order 1^0), described in Section 8.13, give rather direct and accurate information on the parameters of the early universe as summarized in Section 5.5.

5.13 The eras of matter–radiation equality

First we note that the decoupling time for baryonic matter and radiation estimated from (5.49) is about 10^{13} s, or from the results in Table 5.2 and (5.75) (since the universe was matter dominated after $t = t_{\text{dec}}$):

$$t_{\text{dec}} = \frac{t_0}{(1 + z_{\text{dec}})^{3/2}} = 3.7 \times 10^5 \text{ year} \qquad (5.76)$$

for $z_{\text{dec}} = 1100$. It turns out that the energy density of (baryonic) matter, varying as T^{-3}, became equal to that of the (photon) radiation, varying as T^{-4}, at a redshift not very different from that for decoupling. In fact there are several possible eras regarding the relative magnitudes of the energy densities of radiation and of non-relativistic matter. Denoting the photon energy density parameter by $\Omega_r(0)$ as above, *baryon–photon* equality will occur for a redshift given by

$$\frac{\Omega_b(t)}{\Omega_r(t)} = \left[\frac{\Omega_b(0)}{\Omega_r(0)}\right]\left[\frac{R(t)}{R(0)}\right] = \left[\frac{\Omega_b(0)/\Omega_r(0)}{1+z}\right] = 1$$

or

$$\text{(1+z)} = \frac{\Omega_b(0)}{\Omega_r(0)} = \frac{0.042}{4.84 \times 10^{-5}} = 870, \qquad (5.77\text{a})$$

while that for *matter–photon* equality will be at

$$(1 + z) = \frac{\Omega_m(0)}{\Omega_r(0)} = \frac{0.24}{4.84 \times 10^{-5}} = 4950, \qquad (5.77\text{b})$$

and that equating the *matter density* to that of *all relativistic particles* (both photons and neutrinos) will be at redshift

$$(1 + z) = \frac{\Omega_m(0)}{1.58\Omega_r(0)} = \frac{0.24}{7.67 \times 10^{-5}} = 3130 \qquad (5.77\text{c})$$

where we have used the ratio of neutrino to photon energy densities, which from our discussion in Section 5.11 is readily seen to be (9/11) $(4/11)^{1/3} = 0.58$.

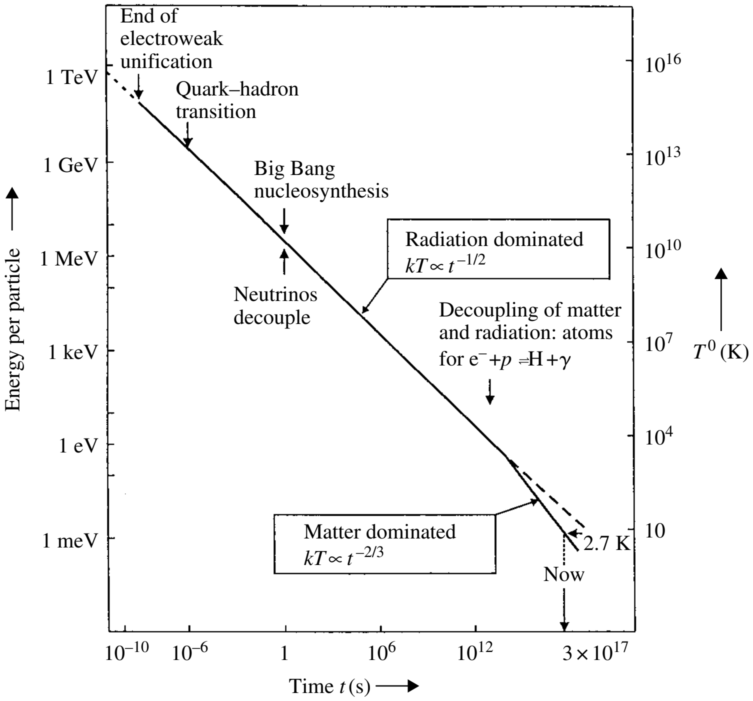


Fig. 5.10 Evolution of the temperature with time in the Big Bang model, with the various eras indicated. See also Fig. 8.2.

These different eras have significance for the various stages in the development of the early universe, as is discussed more fully in Chapter 8. However, we can note here already that when the matter density exceeds that of relativistic particles, that is, for $z < 3000$ as in (5.77c), the gravitational clustering of matter can begin, although it will be opposed by the free streaming away of photons and neutrinos unless it is on very large scales. Dark matter is vitally important here, since the dominance of baryons alone over radiation would not occur until very much later, at $z < 900$, as in (5.77a), and after the decoupling of photons and matter and the formation of atoms. As shown in Chapter 8, without the dominant role of dark matter, it is difficult to see how the observed structures—galaxies, clusters, and superclusters—could have formed so rapidly.

Finally, Fig. 5.10 shows the variation of temperature with time through the radiation and matter eras.

5.14 Summary

- The 'Standard Model' of the universe is based on Einstein's general relativity and the cosmological principle, implying that at early times and on large scales, the universe was isotropic and homogeneous. The 'Big Bang' expansion of the universe follows from Hubble's Law. This expansion is universal and appears the same to all observers, no matter where they are located.

- Hubble's Law describes the linear relation between the redshift z of the light from distant galaxies and the universal expansion parameter R, with $R_0/R_e = \lambda_{\text{observed}}/\lambda_{\text{emitted}} = (1 + z)$.
- The Friedmann equation relates the Hubble expansion parameter $H = \dot{R}/R$ to the total energy density of the universe and the curvature of space (the parameter k).
- The total energy density is the sum of contributions from matter, radiation, and vacuum energy (or so-called dark energy). The vacuum term plays the role of Einstein's cosmological constant.
- The age of the universe is about 14 Gyr. Independent estimates, from radioactive isotope ratios, from stellar population analysis, and from the most recently measured cosmological parameters ($\Omega_m = 0.26$, $\Omega_\Lambda = 0.74$) are all in agreement on the age. The radius of the observable universe, that is, the distance to the present optical horizon, is about 14 Gpc.
- The curvature parameter of the universe, as measured in experiments to be described in Chapter 8, is $k \approx 0$. On largeer scales, the universe is flat, and its total energy—mass energy, kinetic energy, and potential energy—is near to zero.
- The all-pervading and isotropic microwave background radiation has a black body spectrum of $T = 2.725$ K and is the cooled remnant of the Hot Big Bang.
- The baryonic matter and (photonic) radiation energy densities were equal at redshift $z \sim 10^3$, when the universe was about 400,000 years old. Around that time, radiation and matter decoupled, and atoms (mainly hydrogen) started to form because of the recombination of electrons and protons.
- The total matter energy density (including dark matter) started to exceed that of all radiation (both photons and relativistic neutrinos) below a somewhat larger (and earlier) redshift of $z \sim 3000$, and this was an important factor in the early gravitational clustering of matter on the largest scales.

Problems

For all constants required refer to Appendix A. More challenging problems are marked by an asterisk.

(5.1) Assuming that the age of the universe is 14 Gyr and that the total density is equal to the critical density $\rho_c = 9 \times 10^{-27}$ kg m^{-3}, estimate the gravitational binding energy and compare it with the total mass energy of the universe.

(5.2) Calculate the (non-relativistic) escape velocity v of a particle from the surface of a sphere of radius r and uniform mass density ρ. Show that if one assumes Hubble's Law $v = Hr$, the particle will escape provided that $\rho < 3H^2/8\pi\text{G}$.

(5.3) Free non-relativistic fermions of rest-mass m in thermal equilibrium at temperature T are described by the FD distribution (5.56). If $kT \ll mc^2$, show that the number density of particles is $g\left(2\pi mkT/h^2\right)^{3/2}\exp\left(-mc^2/kT\right)$ where g is the number of spin substates.

(5.4) It is estimated that dark vacuum energy today contributes approximately 0.75 to the closure parameter Ω. At what value of the redshift parameter and at what age of the universe would vacuum energy have been less than 10^{-4} of the energy density of radiation?

(5.5) The total amount of energy incident on the Earth's atmosphere from the Sun is 0.135 Joules per cm^2 per second (the solar constant). The Earth–Sun distance is $D = 1.5 \times 10^{11}$ m and the solar radius is $R_S = 7 \times 10^8$ m. Assuming that the Sun is a black body, calculate its surface temperature.

(5.6) It was once proposed that the expansion of the universe could be attributed to an electrostatic repulsion between atoms, on the grounds that the arithmetic values of the electric charges of the electron and the proton might have a very small fractional difference ε. What value of ε would have been necessary? (*Note*: This hypothesis was quickly disproved by experiment, showing that ε was less than 1% of the value required. See proposal by Bondi and Littleton (1959) and experimental disproof by Cranshaw and Hillas (1959).

*(5.7) Find expressions for the dependence of the time t on the density ρ for an expanding 'flat' universe ($k = 0$) dominated (a) by radiation and (b) by non-relativistic conserved matter. Show that, in either case, t is of the same order of magnitude as the time for the gravitational free-fall collapse of a body of density ρ from rest.

*(5.8) In a flat matter-dominated universe ($k = 0$) of age t_0, light from a certain galaxy exhibits a redshift $z = 0.95$. How long has it taken the light signal to reach us from this galaxy? (For a hint, consult equation (8.1)).

(5.9) What is the minimum value of Ω_Λ which will result in an expansion for the case of a flat universe? Neglect the contribution to the energy density from radiation.

(5.10) Prove the statement that the gravitational field anywhere inside a spherical shell of uniform density is zero; and that the field outside a spherical distribution of total mass M is equal to that of a point mass M placed at the centre of the sphere (This is called Newton's Law of Spheres in classical mechanics. In general relativity it is known as Birkhoff's theorem.)

*(5.11) Derive a formula for the age of an open universe ($\Omega < 1$) with zero cosmological constant and negligible radiation density. Use Equation (5.39) and Example 5.3 as a guide. Calculate the age for $\Omega = 0.24$. (*Hint*: Make the substitution $\tan^2\theta = (1+z)\Omega/(1-\Omega)$.

*(5.12) Show that, as indicated above, the temperature below which neutrinos decouple from other matter and radiation in the early, radiation-dominated universe, is $kT \sim 3$ MeV. Hints for the stages in solving this problem are as follows:

(1) Start with the cross-section for the process $e^+ + e^- \rightarrow \nu_e + \bar{\nu}_e$, *via* W-exchange, which is given in Example 3.2 as $\sigma = G_F^2 s/(6\pi)$, where s is the CMS energy squared and it is assumed that masses can be neglected (i.e. $\sqrt{s} \gg m_e c^2$). Evaluate this cross-section in cm^2 when s is in MeV^2.

(2) Show that, treating the electrons and positrons as a Fermi gas of relativistic particles the mean value of s at temperature T is given by $\langle s \rangle = 2\langle E \rangle^2$ where $\langle E \rangle$ is the mean energy of the particles in the distribution, which can be found from equation (5.56).

(3) Calculate the density N_e of electrons and positrons as a function of kT, and hence the rate for the above reaction, $W = \langle \sigma v \rangle N_e$ where v is the relative velocity of the particles, as a function of kT.

(4) Using (5.49) and (5.58) calculate the time t of the expansion as a function of kT, and setting this equal to $1/W$, deduce the value of kT at neutrino decoupling.

6 Nucleosynthesis and baryogenesis

6.1 Primordial nucleosynthesis

In continuation of this discussion of the early universe, we next turn our attention to the synthesis of the nuclei of the light elements—^{4}He, ^{2}H, ^{3}He, and ^{7}Li. The agreement between the predicted and measured abundances of these elements provided early support for the Big Bang hypothesis.

As discussed in Section 5.10, once the universe had cooled to a temperature $kT < 100\,\text{MeV}$, or after a time $t > 10^{-4}$ s, essentially all the hadrons, with the sole exception of neutrons and protons and their antiparticles, would have disappeared by decay. The nucleons and antinucleons would have been present in equal numbers and have nearly, but not quite completely, annihilated to radiation. As described in the next section, once the temperature had fallen below $kT = 20\,\text{MeV}$, a tiny residue of about one billionth of the original numbers of protons and neutrons must have survived to form the stuff of the material universe we inhabit today. The relative numbers of these surviving protons and neutrons would have been determined by the weak reactions

$$\nu_e + \text{n} \leftrightarrow \text{e}^- + \text{p} \tag{6.1}$$

$$\bar{\nu}_e + \text{p} \leftrightarrow \text{e}^+ + \text{n} \tag{6.2}$$

$$\text{n} \rightarrow \text{p} + \text{e}^- + \bar{\nu}_e \tag{6.3}$$

Since at the temperatures considered, the nucleons are non-relativistic, then just as in the analysis of Section 5.12, the equilibrium ratio of neutrons to protons will be governed by the ratio of the Boltzmann factors, so that

$$\frac{N_\text{n}}{N_\text{p}} = \exp\left(\frac{-Q}{kT}\right); \quad Q = \left(M_\text{n} - M_\text{p}\right)c^2 = 1.293 \text{ MeV} \tag{6.4}$$

The rate or width Γ for the first two reactions (6.1) and (6.2) must vary as T^5 purely on dimensional grounds. The Fermi constant G_F from (1.9) or Table 1.5 has dimensions E^{-2}, so the cross-section σ (dimension E^{-2}) must vary as $G_\text{F}^2 T^2$ and the incident flux ϕ, proportional to the neutrino density, as T^3. Hence the width $\Gamma = \sigma\phi$ gets a T^5 factor. On the other hand the expansion rate of the radiation-dominated universe is $H \sim g^{*^{1/2}} T^2$ from (5.59). Hence $\Gamma/H \sim T^3/(g^*)^{1/2}$ and as the universe expands and the temperature falls, the above reactions will go out of equilibrium when $W/H < 1$, where $W = \Gamma/\hbar$. In fact, as described in Chapter 5, at $kT < 3\,\text{MeV}$ neutrinos are already going out of equilibrium with electrons in the process $\text{e}^+ + \text{e}^- \leftrightarrow \nu + \bar{\nu}$, since this has an even smaller cross-section than (6.2) because of the smaller target mass.

Inserting typical values for the cross-sections for (6.1) and (6.2), the neutrino density (5.57) and the Hubble parameter (5.59), it is easy to demonstrate that the freeze-out temperature will be of order 1 MeV—see Problem 6.2. In fact the result of a full calculation for the above reactions gives for the freeze-out a value $kT = 0.80\,\text{MeV}$, so that the initial value of the neutron–proton ratio will be

$$\frac{N_\text{n}(0)}{N_\text{p}(0)} = \exp\left(\frac{-Q}{kT}\right) = 0.20 \tag{6.5}$$

After some time, neutrons will disappear by decay in reaction (6.3). At time t after decoupling, there will then be $N_\text{n}(0)\exp(-t/\tau)$ neutrons and $[N_\text{p}(0) + N_\text{n}(0)\{1 - \exp(-t/\tau)\}]$ protons, with a neutron–proton ratio of

$$\frac{N_\text{n}(t)}{N_\text{p}(t)} = \frac{0.20\exp(-t/\tau)}{[1.20 - 0.20\exp(-t/\tau)]} \tag{6.6}$$

where $\tau = 885.7 \pm 0.8\,\text{s}$ is the presently quoted value of the free neutron lifetime. If nothing else were to happen at this juncture, the neutrons would simply die away by decay and the early universe would consist exclusively of protons and electrons. However, as soon as neutrons appear, nucleosynthesis can begin in a first stage, with the formation of deuterons:

$$\text{n} + \text{p} \leftrightarrow {}^2\text{H} + \gamma + Q \tag{6.7}$$

where the deuteron binding energy $Q = 2.22\,\text{MeV}$. This is an electromagnetic process with a cross-section of 0.1 mb, very much larger than those of the weak processes (6.1)–(6.3), and consequently it stays in thermal equilibrium for very much longer. As indicated below, there is at this time a billion-fold preponderance of photons over nucleons, and the deuterons are not frozen-out until the temperature falls to about $Q/40$, that is, $kT = 0.05 - 0.06\,\text{MeV}$ (see Problem 6.3). As soon as the reverse process of photodisintegration of the deuteron ceases, competing reactions in a second stage leading to helium production take over, for example,

$$\begin{aligned}
{}^2\text{H} + \text{n} &\rightarrow {}^3\text{H} + \gamma \\
{}^2\text{H} + \text{H} &\rightarrow {}^3\text{He} + \gamma \\
{}^2\text{H} + {}^2\text{H} &\rightarrow {}^4\text{He} + \gamma \\
{}^3\text{H} + {}^2\text{H} &\rightarrow {}^4\text{He} + n \\
{}^3\text{H} + \text{p} &\rightarrow {}^4\text{He} + \gamma
\end{aligned}$$

which then lead in a third stage to lithium and beryllium production:

$$\begin{aligned}
{}^3\text{He} + {}^4\text{He} &\rightarrow {}^7\text{Be} + \gamma \\
{}^7\text{Be} + \text{n} &\rightarrow {}^7\text{Li} + \text{p}
\end{aligned}$$

For $kT = 0.05\,\text{MeV}$, corresponding to an expansion time from (5.49) of $t \sim 300\,\text{s}$ for $N_\nu = 3$, the neutron–proton ratio from (6.6) becomes

$$r = \frac{N_\text{n}}{N_\text{p}} = 0.135 \tag{6.8}$$

The expected helium mass fraction, with the mass of the helium nucleus set equal to 4 times that of the proton is then given by

$$Y = \frac{4N_{\mathrm{He}}}{(4N_{\mathrm{He}} + N_{\mathrm{H}})} = \frac{2r}{(1+r)} \approx 0.24 \tag{6.9}$$

The mass fraction Y has been measured in a variety of celestial sites, including stellar atmospheres, planetary nebulae, globular clusters, gas clouds, and so on, with values in the range

$$Y = 0.238 \pm 0.006 \tag{6.10}$$

Problems in evaluating both the predicted and measured values mean that agreement between theory (6.9) and observation (6.10) is still uncertain at the 5% level. Nevertheless, this level of agreement was an early and very important success for the Big Bang model. It should be pointed out here that the observed helium mass fraction is far greater than that which could have been produced in hydrogen burning in main sequence stars; their contribution adds only 0.01 to the ratio Y (see Problem 6.4).

An important feature of nucleosynthesis in the Big Bang scenario is that it accounts not only for ^{4}He but also for the light elements D, ^{3}He, and ^{7}Li, which occur in small but significant amounts, far more in fact than would have survived if they had only been produced in thermonuclear interactions in stellar interiors. The lithium and deuterium abundances give

$$\frac{\mathrm{Li}}{\mathrm{H}} = (1.23 \pm 0.01) \times 10^{-10} \tag{6.11}$$

$$\frac{\mathrm{D}}{\mathrm{H}} = (2.6 \pm 0.4) \times 10^{-5} \tag{6.12}$$

The curves in Fig. 6.1 shows the abundances expected from primordial nucleosynthesis, calculated on the basis of the cross-sections involved, and plotted in terms of the (present day) baryon to photon density ratio. The result (6.12) on the deuterium–hydrogen ratio leads to a value of the baryon density in the range

$$\rho_{\mathrm{B}} = (4.0 \pm 0.4) \times 10^{-28}\ \mathrm{kg\,m^{-3}} \tag{6.13}$$

and a contribution to the closure parameter

$$\Omega_{\mathrm{B}} = 0.044 \pm 0.005 \tag{6.14}$$

corresponding to a number density of baryons $N_{\mathrm{B}} = 0.24 \pm 0.03\,\mathrm{m}^{-3}$. Comparing with the number density of microwave photons (5.52), this yields for the baryon–photon ratio

$$\frac{N_{\mathrm{B}}}{N_\gamma} \approx \frac{\left(N_{\mathrm{B}} - N_{\overline{\mathrm{B}}}\right)}{N_\gamma} = (6.1 \pm 0.6) \times 10^{-10} \tag{6.15}$$

A slightly different value of $(6.5 \pm 0.4) \times 10^{-10}$ is found from the analysis of microwave anisotropies by the WMAP (Wilkinson Microwave Anisotropy Probe), described in Chapter 8. This value for the baryon–photon ratio would imply for the helium fraction, $Y = 0.248$, about 5% larger than the observed value in (6.10).

So, while in the first nanoseconds of the Big Bang, the relative numbers of baryons, antibaryons, and photons would have been comparable (differing only in spin multiplicity factors), most of the nucleons and antinucleons must have later disappeared by mutual annihilation, leaving a tiny—one part per billion—excess of nucleons as the matter of the everyday world, as discussed in the following section.

After the formation of ^{4}He, there is a bottleneck to further nucleosynthesis, since there are no stable nuclei with $A = 5$, 6, or 8. Formation of ^{12}C via the triple-alpha process, for example, is not possible because of the Coulomb barrier suppression, and this has to await the onset of helium burning in stars at high temperatures. Production of heavier elements in stellar fusion reactions at high temperature is discussed in Chapter 10.

It is of interest to remark here that the expected value of the helium mass fraction depends on the assumed number of neutrino flavours N_ν since the expansion timescale to reach a particular temperature as described by (5.59) varies inversely as $\sqrt{g^*}$, the square root of the number of fundamental boson and fermion degrees of freedom. Thus increasing N_ν increases g^*, decreases the timescale, and hence raises the freeze-out temperature T_F determined by the condition $W/H \sim 1$. This leads, through (6.5) to a higher initial neutron–proton ratio and a higher helium mass fraction. Originally, before experiments at the LEP e^+e^- collider, demonstrating that $N_\nu = 3$ (see Fig. 1.13), this argument was used to set a limit on the number of flavours, whereas now it is used to set a better value on the helium mass fraction and the baryon to photon ratio (see Problem 6.1).

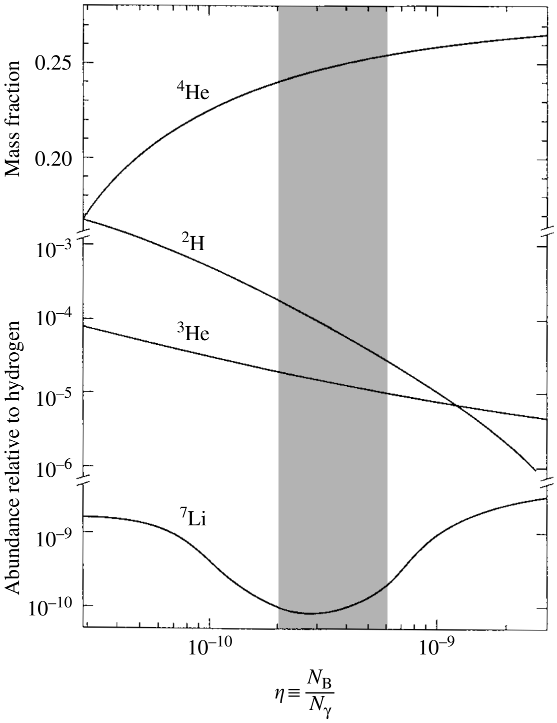


Fig. 6.1 The primordial abundances expected in Big Bang nucleosynthesis of the light elements ^{2}H, ^{3}He, and ^{7}Li, and the mass abundance of ^{4}He, in all cases relative to hydrogen and plotted as a function of the ratio of baryons to photons. The presently observed values of the abundances are given in (6.10)–(6.12). They point to a value of the baryon to photon ratio given in (6.15), having an accuracy of order 10%. The shaded area in the graph shows for comparison the less well-determined value of the ratio—$(4 \pm 2) \times 10^{-10}$—as of 10 years ago (from Schramm and Turner 1998).

6.2 Baryogenesis and the matter–antimatter asymmetry in the Big Bang

One of the most striking features of our universe is the absence of antimatter, although the conservation rules described in Chapter 3 seem to indicate an almost exact symmetry between matter and antimatter. (Recall that the CP asymmetry, observed only in the weak interactions, is measured to be very small.)

Figure 6.2 shows an example, in nuclear emulsion, of the annihilation of an antiproton produced at the Bevatron accelerator at the Lawrence Berkeley Laboratory, where the first antiprotons were observed in 1955. The antiproton annihilates with a nucleon in the emulsion, and four charged mesons, with total energy 1.4 GeV, are produced. Neutral pions will account for the remainder of the annihilation energy of $2Mc^2$, where M is the nucleon mass. We know there is a paucity of antimatter in our own galaxy or in the local cluster, because the primary cosmic ray nuclei, which have been brewed up in stellar reactions over billions of years, and have typically been circulating in the galactic magnetic fields for several million years, are invariably found to be nuclei rather than antinuclei. As an example, Fig. 6.3 shows a case of a chromium nucleus coming to rest in nuclear emulsion carried on a high altitude balloon. Had this been an

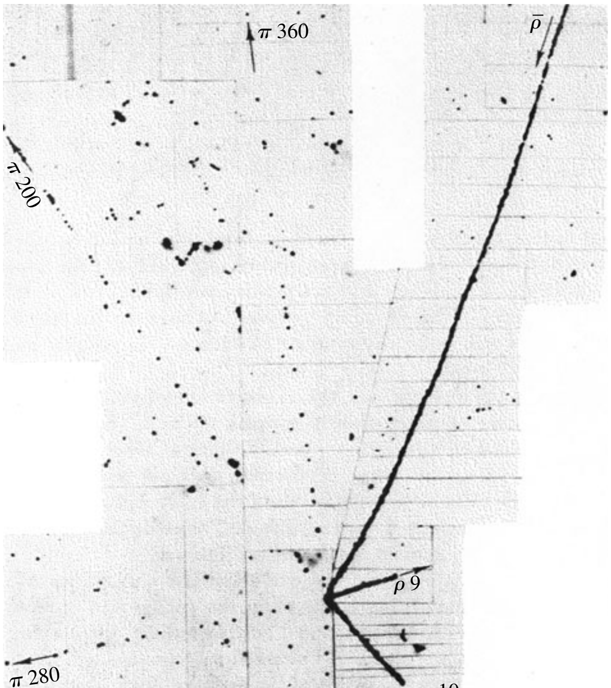


Fig. 6.2 Example of antiproton annihilation in nuclear emulsion. The emulsion consists of a suspension of microcrystals of silver bromide or iodide (of order 0.25μm in radius) in gelatine. Charged particles ionize the atoms they traverse and the electrons liberated form a latent image as they are trapped in the microcrystals. Upon processing, the unaffected halide is dissolved out and the crystals with latent images are reduced to black metallic silver, so forming the tracks, which can be viewed under a microscope. The thick track entering the picture at 1 o'clock is produced by the antiproton as it slows down and comes to rest. The lightly ionized tracks are due to four relativistic charged pions (energies labelled in MeV) which are created in the annihilation process. Together with (unobserved) neutral pions, they account for the total annihilation energy of $2Mc^2$, where M is the nucleon mass. They eject two low-energy protons, at 2.30 and 4 o'clock, from the struck nucleus.

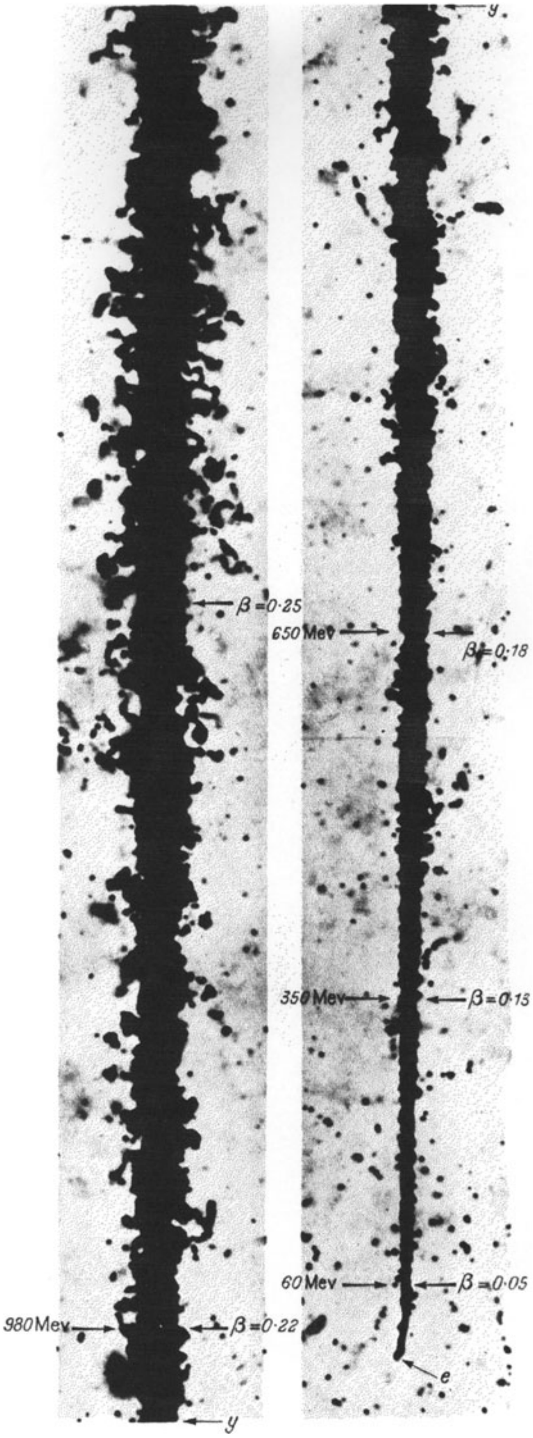


Fig. 6.3 Track of a primary cosmic ray chromium nucleus ($Z = 24$) observed in nuclear emulsion flown on a balloon. The track, of total length 400μm, is shown in two adjacent sections, starting from left top, and terminating at bottom right. As the ionization is proportional to Z^2, the track is initially very dense in comparison with that of the singly charged antiproton in Fig. 6.2. As the nucleus slows down, its velocity becomes comparable with that of electrons in the chromium atom, so it successively collects electrons into the various shells K, L, and so on, the track tapers down and finally it comes to rest as a chromium atom. Had this been an antinucleus of order 100 secondary pions would have been produced as the antinucleus slowed down and annihilated.

antinucleus, the subsequent annihilation would have been 24 times as violent as that in Fig. 6.2.

On a wider scale, there is absolutely no evidence for the intense γ-ray and X-ray emission which would follow annihilation of matter in distant galaxies with clouds of antimatter. The results of nucleon–antinucleon annihilation after recombination of hydrogen atoms at $z \sim 1100$ would also have had significant effects on the Planck black body distribution of the cosmic microwave spectrum. Very low fluxes of positrons and antiprotons do exist in the cosmic rays incident on the Earth's atmosphere, but these can be accounted for in terms of the processes of electron–positron or proton–antiproton pair creation resulting from collisions of high-energy γ-rays or nuclei with interstellar matter.

While it is true that in all laboratory experiments to date, baryon number is conserved, there are in fact no compelling theoretical reasons why this should be so, especially since the early universe would have involved very high temperatures and energies and possibly new types of particles and interactions, which are out of reach at laboratory energies. Indeed, the history of particle physics contains several long-respected conservation laws, which have finally fallen by the wayside, as already described in Chapter 3.

The electric charge, we know, is strictly conserved in all situations, because it is protected by gauge invariance and the existence of a long-range electromagnetic field coupled to the electric charge. An absolute conservation law for baryons ought to have as a consequence the existence of a new long-range field coupled to baryon number. There is no evidence for any such field. If we assume the validity of the equivalence principle described in Chapter 2, the results of the torsion balance experiments in Section 2.3 allow one to set limits for any such field, as follows.

Two bodies of different materials but the same mass will have slightly different baryon numbers—typically by 0.1%—because of variations in binding energy per nucleon with the mass number A, and because of the neutron–proton mass difference and variations with A of the neutron–proton ratio (see Fig. 10.1). So, assuming the validity of the equivalence principle, these experiments find no evidence of any such new long-range field coupled to baryon number and set a limit for such a coupling of less than $10^{-9}G$ where G is the Newtonian constant. These statements also apply to lepton number conservation—again, there are no compelling theoretical reasons for it.

6.3 The baryon–photon ratio in the Big Bang

What predictions does the Big Bang model make for the baryon/antibaryon and the baryon/photon ratios, if we assume strict conservation of baryon number? In the early stages of the Big Bang, when the thermal energy per particle kT (where k is Boltzmann's constant) was large compared with the hadron masses, it is expected that many types of hadrons, including protons and neutrons and their antiparticles, would have been in thermal equilibrium with radiation, being created and annihilated in reversible reactions such as

$$\mathrm{p} + \bar{\mathrm{p}} \leftrightarrow \gamma + \gamma \tag{6.16}$$

Assuming a net initial baryon number of zero, the number density of nucleons and antinucleons at temperature T would be given by (5.56) with $g_f = 2$:

$$N_{\mathrm{B}} = N_{\overline{\mathrm{B}}} = \frac{(kT)^3}{\pi^2 (\hbar c)^3} \int \frac{\left(pc/kT\right)^2 \mathrm{d}\left(pc/kT\right)}{\left[\exp\left(E/kT\right) + 1\right]} \tag{6.17}$$

where p is the three-momentum, m the nucleon mass, and E is the total energy given by $E^2 = p^2c^2 + m^2c^4$. This may be compared with the number of photons in (5.52):

$$N_\gamma = \frac{2.404\ (kT)^3}{\pi^2 (\hbar c)^3} \tag{6.18}$$

The baryons, antibaryons, and photons are in thermal equilibrium and will stay in equilibrium as long as the rate for the back reaction in (6.16) exceeds the universal expansion rate, given by the Hubble parameter H. Eventually, as the expansion proceeds and the temperature falls, the part of the high-energy tail of the photon distribution, with photons above threshold for nucleon–antinucleon pair creation, will become so small that the rate of creation of fresh pairs falls below the expansion rate. Photons cannot produce enough nucleon pairs, nor can nucleons find enough antinucleons with which to annihilate, and the residue of baryons and antibaryons is ‘frozen-out’. The critical temperature at which this occurs depends on the baryon density (6.13), on the nucleon–antinucleon annihilation cross-section and its dependence on velocity, and on the expansion rate. This is a straightforward calculation which is described in Example 6.1 below. Given these parameters, one can solve numerically for the temperature and nucleon density at freeze-out. The predicted result from

Example 6.1 *Calculate the residual baryon–antibaryon ratio after annihilation processes have been completed, assuming baryon–antibaryon equality initially and conservation of baryon number.*

Under these assumptions, the number density of baryons and antibaryons can be easily found by integrating (6.17), assuming that the thermal energy per particle $kT \ll Mc^2$, where M is the proton (or antiproton) mass. This density is then (with $g_f = 2$ for the number of spin substates)

$$N = g_f \left(\frac{2\pi MkT}{h^2}\right)^{3/2} \exp\left(\frac{-Mc^2}{kT}\right) \tag{6.19}$$

The annihilation rate per baryon is $W = <\sigma N v>$, where σ is the proton–antiproton annihilation cross-section at relative velocity v. At the kinetic energies we are concerned with here, up to a few tens of MeV, we can take for $\langle \sigma v/c\rangle$ a value of 80 mb—the figure is not critical. The expansion rate of the universe, which at this point is dominated by radiation, is given from (5.59) as

$$H = 1.66\sqrt{g^*}\,(kT)^2 \cdot \frac{2\pi}{\left(M_{\mathrm{PL}}\hbar c^2\right)} \tag{6.20}$$

where the total number of substates $g* \sim 10$ as in Section 5.10, and $M_{\mathrm{PL}}c^2 = 1.2 \times 10^{19}$ GeV is the Planck energy. Thus, inserting the values of the constants, one obtains for the ratio

$$\frac{H}{W} = 3.2 \times 10^{-19} \frac{\left[\exp\left(Mc^2/kT\right)\right]}{\left[Mc^2/kT\right]^{1/2}} \tag{6.21}$$

This ratio varies from 43.0 at $kT = 19$ MeV, to 3.7 at $kT = 20$ MeV, to 0.41 at $kT = 21$ MeV, that is, by a factor of 10 for every change in kT of 1 MeV. The freeze-out temperature, when $H/W \sim 1$, is clearly close to $kT = 20$ MeV, and the corresponding ratio of baryon to photon densities is, using (6.17)–(6.19):

$$\frac{N_\mathrm{B}}{N_\gamma} = \frac{N_{\overline{\mathrm{B}}}}{N_\gamma} = 0.72 \times 10^{-18} \tag{6.22}$$

the above example is

$$kT\ (\text{critical}) \approx 20\ \text{MeV}; \quad \frac{N_\mathrm{B}}{N_\gamma} = \frac{N_{\overline{\mathrm{B}}}}{N_\gamma} \sim 10^{-18} \tag{6.23}$$

Thus the annihilation of nucleons and antinucleons is almost but not quite complete. Simply because the universe is expanding, there remains a tiny residue of baryons and antibaryons. Subsequent to this freeze-out stage, there would be no further nucleon–antinucleon annihilation or creation and the above ratios should hold today.

In contrast, the observed value of the baryon–photon ratio shown in (6.15), is much larger. To summarize, the observed ratios to be compared with the predictions in (6.23) are

$$\frac{N_\mathrm{B}}{N_\gamma} \approx 10^{-9} \quad \frac{N_{\overline{\mathrm{B}}}}{N_\mathrm{B}} < 10^{-4} \tag{6.24}$$

So, the Big Bang hypothesis gets the baryon–photon ratio wrong by a factor of 10^9 and the antibaryon–baryon ratio wrong by a factor of at least 10^4. Of course, it is possible to avoid this problem by arbitrarily assigning an initial baryon number to the universe, but this would be quite large ($N_\mathrm{B} \sim 10^{79}$!) and arbitrary, and in any case not possible in the inflationary model described in Chapter 8. It seems more sensible to try to understand the observed values in terms of (hopefully) known physics. This takes us back to a seminal paper in 1967 by Andrei Sakharov who proposed a possible way out.

6.4 The Sakharov criteria

Sakharov pointed out the fundamental conditions necessary to achieve a baryon–antibaryon asymmetry. Assuming a baryon number $B = 0$ initially, a baryon number asymmetry could obviously only develop as a result of baryon number violating reactions, but in addition one would need two further conditions. In fact the three conditions are

- B-violating interactions.
- Non-equilibrium situation.
- CP and C violation.

The first requirement is obvious and its possibility has been discussed in connection with the grand unified theory (GUT) models and the search for proton decay in Section 4.3. At present, there is no direct *laboratory* evidence that baryon number is violated, so we just retain it as an assumption. The second condition follows from the fact that, in thermal equilibrium, the particle density depends only on the particle mass and the temperature. Since particle and antiparticle have identical masses by the CPT theorem (see Section 3.14), no asymmetry could develop. Put another way, at equilibrium, any reaction which destroys baryon number will be exactly counterbalanced by the inverse reaction which creates it. Third, as pointed out in Chapter 3, C and CP violation are necessary if antimatter is to be distinguished unambiguously from matter on a cosmic scale. We now discuss models for which these conditions can hopefully be met.

6.5 The baryon–antibaryon asymmetry: possible scenarios

The precise mechanism responsible for generating a baryon–antibaryon asymmetry of the observed magnitude is presently unknown, although several models have been suggested over the last two or three decades. Three of the possibilities are

- GUT baryogenesis.
- Electroweak baryogenesis.
- Baryogenesis *via* leptogenesis.

6.5.1 Baryogenesis in the SU(5) GUT model

We first consider baryon number violation via the SU(5) model of grand unification. In that model, already discussed in Chapter 4, quarks and leptons are incorporated in the same multiplets, and quark–lepton transitions can therefore take place, with the interesting possibility of proton decay. Indeed this was one of the predictions in Sakharov's 1967 paper. However, he assumed the mediating bosons of the unified theory had masses of order the Planck mass, resulting in prediction of an unobservably long lifetime of 10^{50} years. For example, as shown in Section 4.3, in the SU(5) model a proton can transform into a pion and a positron *via* virtual X-boson exchange. In this transition, both the baryon number B and the lepton number L have decreased by one unit, so that the difference $(B-L)$ is conserved, while $(B+L)$ is violated. This turns out to be a crucial feature in the discussion of possible mechanisms for generating baryon asymmetry.

The mediating 'leptoquark' bosons X, Y, and their antiparticles of the GUT symmetry are supposedly created in the Big Bang on a 10^{-40} s timescale, and are expected to decay out of thermal equilibrium. The requirements are for two decay channels, say 1 and 2, of different baryon number. Suppose that x and $(1-x)$ are the branching ratios for decay of X to modes with baryon number B_1 and B_2 respectively. For the antiparticle $\overline{X}$, let the ratios be $\bar{x}$ and $(1-\bar{x})$, with baryon numbers $-B_1$ and $-B_2$. Since the numbers of X and $\overline{X}$ particles are equal, by the CPT theorem, discussed in Chapter 3, the net baryon asymmetry per $X\overline{X}$ pair will be

$$A = xB_1 - \bar{x}B_1 + (1-x)\,B_2 - (1-\bar{x})\,B_2 = (x-\bar{x})\,(B_1 - B_2) \qquad (6.25)$$

B violation ensures that $B_1 \neq B_2$, and CP violation that $x \neq \bar{x}$, so that the asymmetry will be non-zero. It is to be noted that C violation alone, with CP conservation, would give an X decay rate at angle θ equal to the $\overline{X}$ decay rate at $(\pi - \theta)$, and therefore the same overall rate when integrated over angle. CP violation is necessary to ensure different partial decay rates for particle and antiparticle in one particular channel (1, for example).

In this model, baryon asymmetry arises from decays of the X and Y 'leptoquark' bosons if their masses are such that out-of-equilibrium decays take place. These decays can certainly generate a baryon asymmetry of the required magnitude ($\sim 10^{-9}$) via non-conservation in the quantity $(B + L)$. Unfortunately, however, it appears that any baryon asymmetry originating in such GUT processes is likely to be washed out by subsequent non-perturbative processes (instantons) associated with the electroweak interaction, as described below. This is true for the original SU(5) model and for other proposed GUT symmetries, for example the group called SO(10), which is more extensive than SU(5). Most importantly, SO(10) has room for an extra U(1) right-handed Majorana neutrino singlet state, not protected from radiative corrections by the gauge symmetry of the electroweak SU(2) × U(1) group, and whose mass can therefore be arbitrarily large. The possible role of such neutrinos in baryogenesis is discussed below.

6.5.2 Baryogenesis in the electroweak model

Baryogenesis is also possible in principle *via* $(B + L)$ asymmetries occurring in (first-order) phase transitions in the electroweak sector, which will occur at energies of the order of 100 GeV, the mass scale of the gauge bosons W and Z. The sphaleron mechanism generating this asymmetry is described in the next section. The Standard Model also incorporates the required degree of C and CP violation, as described in Chapter 3. However, the out-of-equilibrium condition that is also necessary, occurring during the phase transition, appears to be too feeble to produce the observed asymmetry. Considerable extensions to the Standard Model would be required if electroweak baryogenesis is to work, and this type of model is currently disfavoured.

6.5.3 Baryogenesis via leptogenesis

What is thought to be a more likely scenario, first proposed by Fukugita and Yanagida (1986) is that a *lepton* asymmetry, rather than a baryon asymmetry, is generated first at the GUT energy scale *via* the out-of-equilibrium decay of massive Majorana neutrinos, N. These are postulated as right-handed singlet states, for example, as components of an SO(10) GUT. Majorana neutrino decays do not conserve lepton number. An example of such a decay (by the conventional weak interaction) would be into a light neutrino and Higgs:

$$N \rightarrow H + v \qquad (6.26)$$

Again, to generate a lepton asymmetry, the N particles must have out-of-equilibrium number densities. So the decay rate and width must be less than the

Hubble parameter, $W = \Gamma/\hbar \ll H$. This requirement puts constraints on the N mass. Most importantly, however, the resulting lepton asymmetry conserves the quantity $(B - L)$.

In this model, the lepton asymmetry is subsequently converted into a baryon asymmetry by non-perturbative processes at the lower-energy scale of the electroweak interactions. Here we dip into the somewhat exotic scenario of gauge anomalies (i.e. divergent terms in the axial–vector weak currents), instantons and sphalerons. Instantons are examples of single events in field theory. One can think of such events as comparable to the process of radioactive decay by single alpha particle emission through a potential barrier (discussed in Chapter 10). As we noted in Chapter 3, the vacuum state (i.e. the state of lowest energy) in the electroweak model can be quite complex. Indeed, it turns out that there is an infinite number of degenerate vacuum states with different topologies, that is, different baryon and lepton numbers. Adjacent vacua differ in the value of $(B + L)$ by $2N_f$ where $N_f = 3$ is the number of quark or lepton flavours, and they are separated by potential barriers with a height of the order of the electroweak vacuum expectation value $v \sim 200$ GeV. On the other hand, the quantity $(B - L)$ is 'anomaly-free' and conserved. It turns out that the change in lepton and baryon numbers between adjacent vacua is $\Delta L = \Delta B = 3$. At normal energies, such changes can happen only by quantum-mechanical tunnelling *through* the barrier between one vacuum state and the next. As first shown by 't Hooft in 1973, such a so-called *instanton* process is enormously suppressed by a factor of order $\exp(-2\pi/\alpha_\mathrm{w}) \sim 10^{-86}$, where α_w is the weak coupling.

However, an important observation by Kuzmin *et al.* (1985) was to point out that at high enough temperatures, $kT > v$, thermal transitions can take place by jumping *over* the barrier, *via* a 12-fermion interaction referred to as a *sphaleron* (the name comes from the Greek for an unstable state: the sphaleron is a saddle point in configuration space, which sits on the top of the barrier, and can jump either way). Typical $\Delta B = \Delta L = 3$ transitions would be

$$\begin{aligned}(u + u + d) + (c + c + s) + (t + t + b) &\rightarrow \mathrm{e}^+ + \mu^+ + \tau^+ \\ (u + d + d) + (c + s + s) + (t + b + b) &\rightarrow \bar{\nu}_e + \bar{\nu}_\mu + \bar{\nu}_\tau \qquad (6.27)\end{aligned}$$

In these transitions, three quarks and one lepton of each of the $N_f = 3$ generations are involved. The degree of baryon–antibaryon asymmetry thus generated—typically of the order of half the magnitude of the original lepton asymmetry—depends on the assumed masses M of the massive Majorana neutrinos N. At the same time, the masses of the light neutrinos, which can be estimated from observations of neutrino oscillations described in Chapter 9, are connected with the M values *via* the so-called 'see-saw' mechanism discussed in Chapter 4. The crucial point here is that the $(B - L)$ asymmetry generated by the N particles at the GUT scale is preserved in going through the electroweak transition. This is in contrast with the GUT-generated baryon asymmetry of Section 6.5.1, where the $(B + L)$ asymmetry is washed out by the very same sphaleron processes which convert lepton asymmetries to baryon asymmetries.

It is remarkable that the Majorana masses $\sim 10^{10} - 10^{13}$ GeV required to fit the light neutrino masses by the see-saw formula also seem to give lepton asymmetries of the correct magnitude to provide, in turn, baryon asymmetries of about the magnitude observed in (6.15). Indeed, one can reverse the argument and deduce that from the observed baryon asymmetry ($\sim 10^{-9}$), the above mechanism would only work if the light neutrino masses were in the range 0.01–0.1 eV/c^2, exactly where oscillation experiments place them.

There are, however, some problems with this model. Such massive Majorana neutrinos would have to be created during the 'reheating' phase immediately following inflation, as discussed in Chapter 8. A difficulty here is that, in order to produce massive neutrinos, a high reheating temperature is required, and in supersymmetric theories this could result in an overproduction of gravitinos (the massive spin $3/2$ fermionic partners of the gravitons). The rapid decay of these massive gravitinos would certainly produce hadrons, and unfortunately this has the effect of completely changing the parameters in nucleosynthesis, so that the good agreement between predictions and observations described in Section 6.1 would be lost.

One suggestion for avoiding this problem has been to appeal to the decay of inflatons (the fundamental bosons of the inflation model discussed in Chapter 8) to produce the Majorana particles, $\varphi \to N+N$. Thus, the reheating temperature following the end of inflation can be $kT \ll M$ and the above difficulty can be avoided (Yanagida 2005).

If one believes that all the difficulties can be circumvented, the leptogenesis model seems to provide a reasonably self-consistent picture, although the ideas presented are of course very speculative. We do not even know yet whether neutrinos are Majorana or Dirac particles, and the necessary CP violating phases in the massive neutrino sector are completely unknown. However, the link between the observed small masses of the light neutrinos and the observed baryon–antibaryon asymmetry of the universe, *via* massive Majorana neutrinos and the see-saw mechanism, if it can be confirmed, would be one of the outstanding achievements of particle astrophysics. While the true origin of the universal baryon asymmetry is at present unknown, the situation for the future looks hopeful, with enormous experimental efforts on measurement of neutrino masses and mixings, and the hoped-for discovery of the Higgs boson and of supersymmetry.

6.6 Summary

- The observed abundances of the light elements deuterium (^{2}H), helium (^{3}He and ^{4}He), and lithium (^{7}Li) can be understood by their creation in nucleosynthesis at temperatures $kT \sim 0.1$ MeV in the first minutes following the Big Bang. Together with the microwave background radiation and the redshift, the light element abundances provide very strong support for the Big Bang hypothesis. The baryon density from the synthesis of the light elements in the first minutes of the universe accounts for only a small part of the total matter density: most of the matter is dark matter.
- The observed strong asymmetry between matter and antimatter has to be ascribed to special baryon number violating and CP violating interactions operating at a very early stage of the Big Bang when the temperature was very high.
- There is a realistic prospect that, if heavy Majorana neutrinos exist, the baryon asymmetry of the universe may be directly connected with the

small observed masses of the light neutrinos *via* the see-saw mechanism, coupled with non-perturbative interactions on the electroweak scale.

Problems

More challenging problems are shown by an asterisk

*(6.1) Calculate the expected mass ratio of primordial helium to hydrogen, as in (6.9), but for the case of different numbers of neutrino flavours $N_\nu = 3, 4, 5, 6, \ldots$. Show that each additional flavour will increase the expected ratio by about 5%. Calculate also the expected mass ratio for $N_\nu = 3$ if the neutron—proton mass difference was 1.40 MeV/c^2 instead of 1.29 MeV/c^2, but the free neutron lifetime was unaffected.

*(6.2) In discussing the neutron/proton equilibrium ratio (6.8) it was stated that the rate or width W for the reaction $\nu_e + n \rightarrow e^- + p$ varied as T^5. Verify this directly, referring back to Section (1.8) to compute the cross-section for the above reaction as a function of T and using the relevant flux density to compute W from (1.14). Assume that all particles have kinetic energies kT, such that $m_e c^2 \ll kT \ll M_p c^2$, that is, treat the nucleons as non-relativistic and essentially stationary, while the leptons are extreme-relativistic. Comparing with the expansion rate (5.59), estimate the temperature at which the neutrons and protons 'freeze-out' of equilibrium.

(6.3) Estimate the value of kT below which, in an expanding universe, deuterium will undergo 'freeze-out' from the reaction $n + p \leftrightarrow d + \gamma + Q$, where $Q = 2.22$ MeV is the binding energy of the deuteron. Proceed by first deriving an analytic expression for the fraction of cosmic microwave photons with $E \gg kT$. Assume that the photodisintegration cross-section is $\sigma = 0.1$ mb, and take the Hubble parameter from (5.59), with $g* = 10$.

(6.4) The Sun has a measured luminosity of $3.9 \times 10^{26} W$. It generates its energy from the conversion of hydrogen to helium in thermonuclear fusion reactions, an energy of 26 MeV being liberated for each helium nucleus formed. If the Sun's output has been constant at the above value for 5 Gyr, what is the mass fraction of helium in the Sun?

7 Dark matter and dark energy components

7.1 Preamble

In Chapter 5 we already noted that it appears that a large fraction of the matter in the universe is dark (i.e. non-luminous) matter. The need to postulate such dark matter was noted as early as the 1930s by Zwicky, who observed that galaxies in the Coma cluster seemed to be moving too rapidly to be held together by the gravitational attraction of the visible matter. Obviously, we can hardly be satisfied with our picture of the universe until the nature and distribution of such vast quantities of matter has been settled. For example, an important question is whether this dark matter is in the form of new types of (stable) elementary particle, which have been roaming around since the earliest stages of the Big Bang: and if so, what are such particles, and why have we not met with them in accelerator experiments? Or, could it be that some of the dark matter is agglomerated in the form of non-luminous stellar objects made out of the same matter as ordinary stars, or as mini black holes or whatever?

According to present ideas, the quark and lepton constituents of matter with which we are familiar in experiments at accelerators, produced in the numbers foreseen by the model of nucleosynthesis in the early universe described in Chapter 6, can account for only about 4% of the present energy density of the universe. Dark matter is estimated to account for some 20% of the total energy density, but the bulk of the energy density—that is, some 76%—has to be assigned to 'dark energy', which in Chapter 5 was identified with vacuum energy. However, the true source of the dark energy—like that of the dark matter—is unknown at present.

Before proceeding further, we should recall that, of the total baryonic contribution deduced from primordial nucleosynthesis as described in Chapter 6, only about 10% is accounted for by the luminous matter in stars and galaxies. Hot gas in galaxy clusters and intergalactic hydrogen accounts for a further 40%, leaving half the baryons unaccounted for. As described in Section 7.5, some baryons are located in dark, compact stellar-like objects (MACHOs, or massive astrophysical compact halo objects) in galactic halos, detected by their gravitational lensing of light signals from more distant stars. However, these can account for only a small part of the baryon contribution. Recent observations, discussed in Section 7.7, indicate that the missing baryonic matter may be associated with blazers (see Section 9.14.2).

In this chapter, we first present the evidence for the existence of dark matter, then describe briefly some of the possible dark matter candidates and the attempts to detect them directly. Finally, we describe the evidence for the acceleration of the Hubble expansion from the study of high redshift supernovae, and the consequences for the dark energy/cosmological constant.

7.2 Dark matter in galaxies and clusters

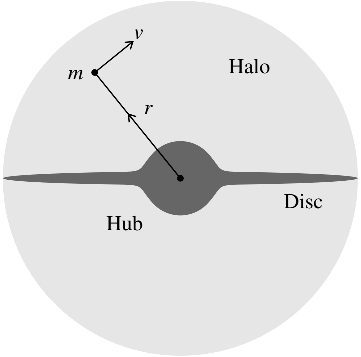


Fig. 7.1 An end-on view of a spiral galaxy, consisting of a central hub, a disc, and a possible halo of dark matter.

The classical evidence for dark matter comes from the measurement of the rotation curves of velocity versus radial distance for stars and gas in spiral galaxies. This has given strong, if indirect, indications for the existence of 'missing' mass, in the form of non-luminous matter. Consider, for example, a star of mass m at distance r from the galactic centre, moving with tangential velocity v as shown in Fig. 7.1. Then equating gravitational and centrifugal forces we obtain

$$\frac{mv^2}{r} = \frac{mM\,(<r)\,\mathrm{G}}{r^2} \tag{7.1}$$

where $M\,(<r)$ is the mass inside radius r. A spiral galaxy such as our own has most of the luminous material concentrated in a central hub plus a thin disc. For a star inside the hub, we expect $M\,(<r) \propto r^3$ and therefore $v \propto r$, while for one located outside the hub, $M \sim$ constant and therefore we expect $v \propto r^{-1/2}$. Hence, the velocity should increase at small r and decrease at large r. On the contrary, for many spiral galaxies, the rotation curves are quite flat at large r values. An example is shown in Fig. 7.2. This has led to the suggestion that the bulk of the galactic mass—typically 80–90%—is in the form of dark matter in a halo as in Fig. 7.1.

Surveys of galaxy clusters show that much of the 'visible' mass is in the form of very hot, X-ray emitting gas. The gas temperature (typically 10^7 to 10^8 K) estimated from the X-rays measured with the ROSAT satellite implied velocities of gas particles far in excess of the escape velocities as deduced from the visible mass. If the gas is bound by gravitational forces, suggested by the fact that it appears concentrated towards the cluster centres, the greater part (at least 80%) of the total mass must be dark matter.

The major surveys of galaxies and galaxy clusters, such as the infrared IRAS satellite survey, comparing the motional energy with the gravitational energy, also provide evidence for dark matter. This analysis is based on the *virial theorem* of classical mechanics. This relates the time average of the potential energy $<V>$ to that of the kinetic energy $<E>$ of a bound system of i non-relativistic particles of masses m_i, velocities $\mathbf{v}_i$, momenta $\mathbf{p}_i$, and kinetic energies E_i, interacting *via* a central inverse square law of force, $\mathbf{F}_i$. The virial is defined as $W = \Sigma \mathbf{p}_i \cdot \mathbf{r}_i$, where $\mathbf{r}_i$ is a position vector measured from some arbitrary origin. On differentiation with respect to time this becomes

$$\begin{aligned}\frac{\mathrm{d}W}{\mathrm{d}t} &= \sum \dot{\mathbf{p}}_i \cdot \mathbf{r}_i + \sum \mathbf{p}_i \cdot \dot{\mathbf{r}}_i = \sum m_i \ddot{\mathbf{r}}_i \cdot \mathbf{r}_i + \sum m_i \left|\mathbf{v}_i\right|^2 \\ &= \sum \mathbf{F}_i \cdot \mathbf{r}_i + 2\sum E_i = \sum \left(\frac{\partial V_i}{\partial r_i}\right) \cdot \mathbf{r}_i + 2\sum E_i \\ &= -\sum V_i + 2\sum E_i\end{aligned}$$

where in the last line we have used the fact that the gravitational potential varies as $1/r$. Averaged over a time interval T, the virial of a *bound* system

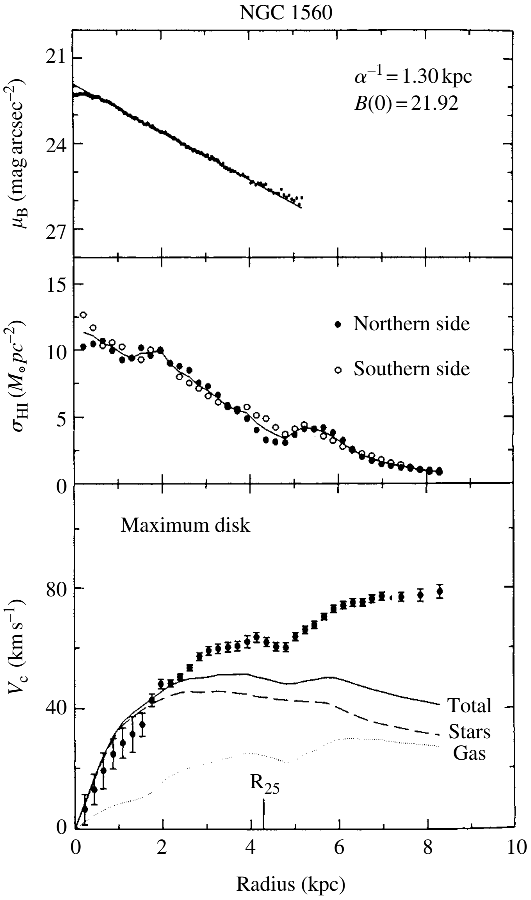


Fig. 7.2 Example of rotation curves for the spiral galaxy NGC 1560. In the top panel the luminosity is plotted against radial distance, showing an exponential fall-off. The middle panel shows the luminosity in the $H\alpha$ line. In the bottom panel, the points show the observed tangential velocities of stars in this galaxy as a function of radial distance. The curves show the expected values obtained by numerical integration of the mass inside a particular radius as in (7.1), with the contributions from stars and gas shown separately. They are clearly unable to account for the observed velocities at large radii (from Broeils 1992).

$\langle W\rangle = (1/T)\int (dW/dt)\,dt \to 0$ as $T \to \infty$, so that the time-averaged $<V>$ $=2<E>$.

Measurements have been made with the Chandra satellite experiment (Allen *et al.* 2004) recording X-rays from large galaxy clusters, which contain many hundreds of galaxies embedded in them. As they are the largest bound systems known, the assumption made is that they represent a fair sample of the material of the entire universe. The clusters contain X-ray emitting gas at temperatures of order 10^6 K, and the virial theorem shows that dark matter is required to hold the clusters together. The X-ray observations actually allow one to estimate the ratio of the mass of hot gas to dark matter in a cluster. Making the reasonable assumption that this ratio is the same for all clusters, one can adjust the distance scale and hence absolute luminosity for each cluster to get the best fit to a universal value of the gas to dark matter ratio. In this way it could also be shown that the early deceleration of the expanding universe, due to the gravitational attraction of matter, was replaced by acceleration about 6 billion years ago. These results agreed perfectly with earlier, independent measurements, that dark energy accounts for 76%, dark matter for 20%, and baryonic matter for 4% of the energy of the universe. These previous observations came from high redshift supernovae, discussed later in this chapter (Section 7.14). It is in fact remarkable and heartening that quite different techniques for estimating the basic parameters of the universe are in such good agreement.

Some of the most remarkable evidence for dark matter comes from the observation of emission lines from very distant clouds of hydrogen, indicating redshifts of $z = 5$ or 6, located in vast galaxies of what appears to be dark matter, accounting for at least 99% of the total mass. Dark matter seems also to be required from quite independent considerations of the level of fluctuations in the cosmic microwave background and the growth of structure in the early universe, as discussed in Chapter 8. These density fluctuations are observed to be of order $\Delta\rho/\rho \sim 10^{-5}$, and fluctuations 2–3 orders of magnitude larger would have been necessary if formation of galaxy and galaxy clusters was to be achieved by gravitational collapse of baryonic matter alone, once it had decoupled from radiation at $z \sim 1000$. On the other hand, as discussed in Section 5.13, the existence of (cold) dark matter with $\Omega_{\text{cdm}} \sim 0.20$ would have led to dominance of matter over radiation at a higher redshift ($z \sim 3000$) and more effective gravitational collapse of matter (with the gravitational field of the dark matter dragging baryonic matter along with it) from a much earlier epoch.

Finally, the masses of galaxy clusters and the contribution of dark matter can be estimated directly by their effects on the images of more distant quasars, due to the process of gravitational lensing, which is discussed in the following sections. It has the advantage that it avoids some of the assumptions necessary for other methods.

7.3 Gravitational lensing

Very important information regarding the amount and location of dark matter has come from gravitational lensing, which we therefore discuss at some length. The gravitational deflection of photons passing by a point mass M at a distance of closest approach b was given in Section 2.6 by the formula (2.28):

$$\alpha = \frac{4\text{G}M}{c^2 b} \tag{7.2}$$

This is the deflection predicted by Einstein's general theory of relativity, being exactly a factor 2 larger than the deflection one obtains according to Newtonian mechanics (see Problem 7.1). The correctness of the Einstein prediction was first demonstrated by the 1919 solar eclipse expedition by Eddington to the island of Principe, which measured the deflection of light from stars close to the Sun's limb.

The gravitational deflection of light implies that massive objects may act as *gravitational lenses*, as foreseen by Einstein even before the relation (7.2) had been tested experimentally. Suppose in Fig. 7.3 that S is a source of light (star) and that the rays to the observer O pass close to a massive point lensing object L of mass M. The diagram represents the situation in the plane defined by O, S, and L, and is the gravitational analogue of a thin lens system in optics. In the general case, the source and lens will not be collinear with the observer and there will then be two images of the source, S1 and S2. Then if α denotes

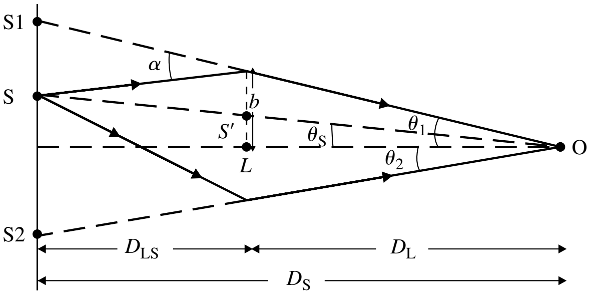


Fig. 7.3 The two images S1 and S2 of a source S formed from gravitational lensing by the point mass *L*.

the gravitational deflection and *b* the closest distance of approach, we have from (7.2):

$$\alpha D_{LS} = D_S\,(\theta_1 - \theta_S)$$

$$\theta_S = \theta_1 - \left(\frac{4GM}{bc^2}\right)\left(\frac{D_{LS}}{D_S}\right) = \theta_1 - \left(\frac{4GM}{c^2}\right)\left(\frac{D_{LS}}{D_S D_L}\right)\left(\frac{1}{\theta_1}\right) \tag{7.3}$$

In the collinear case, $\theta_S = 0$. Then we can write

$$\theta_1 = \theta_E = \left[\left(\frac{4GM}{c^2}\right)\left(\frac{D_{LS}}{D_S D_L}\right)\right]^{1/2} \tag{7.4}$$

where θ_E is the angle of the so-called *Einstein ring*. In this collinear case, the image of S is a ring of light centred on the line of sight. For finite θ_S and a point lensing mass, however, one obtains just two images lying in the plane defined by the source, lens, and observer, with angles which are solutions of the quadratic (7.3):

$$\theta_{1,2} = \left[\theta_S \pm \sqrt{\theta_S^2 + 4\theta_E^2}\right] \tag{7.5}$$

7.4 Evidence for dark matter from gravitational lensing

The above analysis assumed a point lensing mass. Often, the lensing object or objects will be extended in space, and more complex, multiple images are then formed. Examples of lensing were first observed for very intense and very distant sources called quasi-stellar radio sources or *quasars*, which are in fact the most powerful radio and optical sources known (see Section 9.14). Quasars are examples of galaxies with very active galactic nuclei (AGNs), and are almost certainly powered by the gravitational energy from massive black holes (see Section 9.15). In the case of quasars the lensing mass is a ‘foreground’ galaxy or galaxy cluster. An early example of a doubly imaged quasar is shown in Fig. 7.4.

When the lensing galaxies or clusters are extended objects, the lensed images of more distant objects can appear as multiple arcs, as shown in Fig. 7.5. Since in multiply imaged events, the different images involve different light paths, time

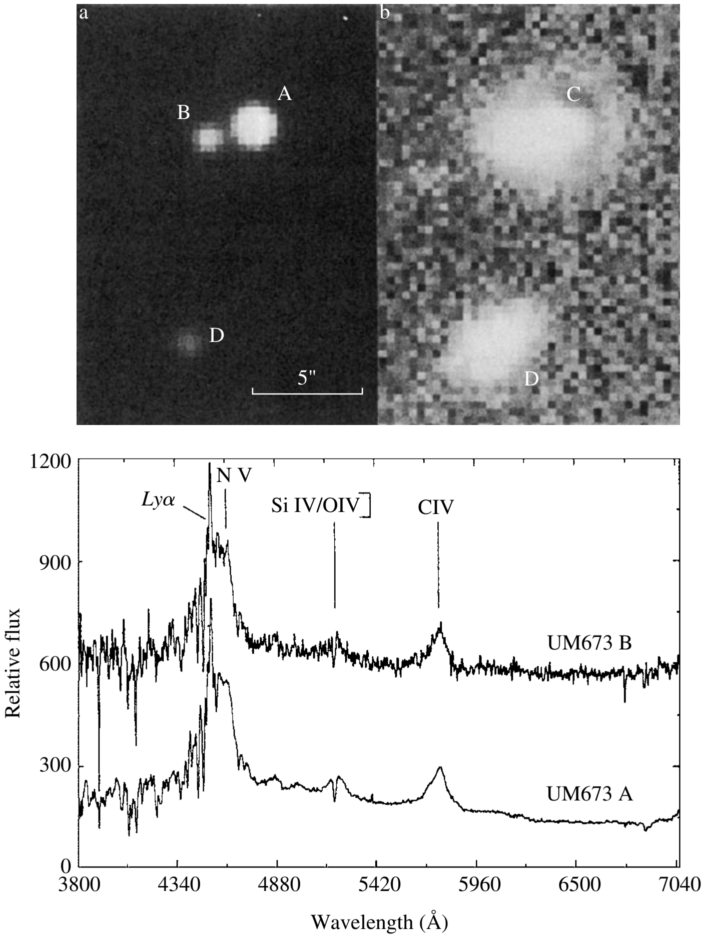


Fig. 7.4 An example of the double image of a quasar, observed by the European Southern Observatory, as it is lensed by a foreground galaxy. The top picture shows (at left) the quasar CCD image split into two parts, A and B. Subtracting these images from the frame reveals (at right) the lensing galaxy marked C. Object D is a background galaxy. The plot of the wavelength response at bottom shows that the two images, separated by 2.2 s of arc, have identical spectra (from Surdej *et al.* 1987).

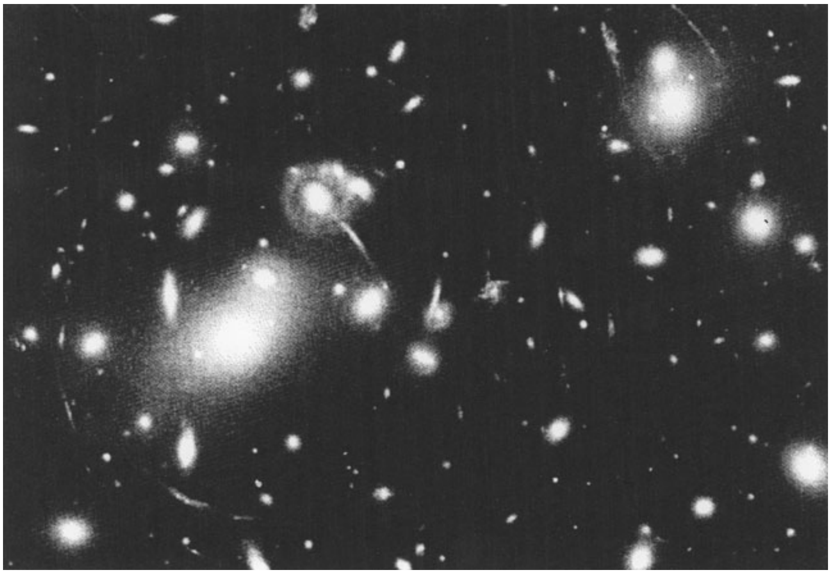


Fig. 7.5 Multiple images, seen in the form of long faint arcs, due to lensing effects by the galaxy cluster Abell 2218 of more distant galaxies. Picture by Hubble Space Telescope (Kneib *et al.* 1996).

delays will be involved. The path lengths are proportional to the distance scale, that is, to the inverse $1/H_0$ of the Hubble parameter, so that study of multiply-imaged quasars offers a method to determine H_0. However, the important thing is that by measurement of the multiple images of such distant quasars, the total gravitating mass of the foreground galaxy or cluster can be measured. The total mass density of the universe found in this way also indicates a value for the closure parameter associated with the matter content of $\Omega_m \approx 0.24$, as quoted in (5.33).

An example of the power of the lensing technique in providing quite compelling evidence for dark matter is provided by observations (Clowe *et al.* 2006) made using a combination of the Hubble Space Telescope, the ESO Very Large Telescope, the Magellan telescope, and the Chandra X-ray satellite. They observed a system of two galaxy clusters apparently having passed through each other. Using gravitational lensing of more distant galaxies by this cluster, it was possible to map the gravitational potential in the cluster, and thus the total matter distribution. On the other hand, the X-ray signals indicate the distribution of hot gas (i.e. the plasma of baryonic matter), and of course the luminous matter in stars is observed by the optical telescopes. The dark matter appeared in two distinctly separated regions. The X-rays were also found to be located in two regions, which are, however, well separated from the dark matter regions (see Fig. 7.6). The importance of these observations is that the regions of dark matter and of baryonic matter are distinct and well separated. In such a collision, as the two clusters pass through each other, the gas clouds would be slowed through electromagnetic interactions, but not the dark matter clouds, presumably subject only to the weak and gravitational interactions. Because of this spatial separation of dark matter and of baryonic matter, the observations cannot be explained as an artefact, due, for example, to a modification of the law of Newtonian gravity at large distances.

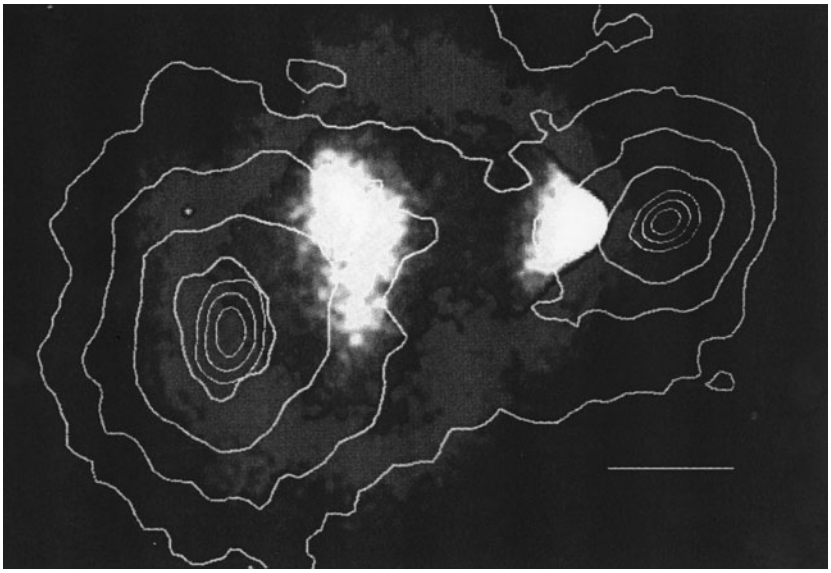


Fig. 7.6 The galaxy cluster 1EO657-558 observed by Clowe *et al*. (2006), interpreted as a case of two clusters passing through each other. The two white areas show the sources of X-rays measured by the Chandra satellite, and correspond to the regions of hot plasma (baryonic matter). The contour lines indicate the regions of dark matter deduced from the gravitational lensing of background galaxies (observed with optical telescopes) and these are seen to be distinctly separated from the plasma regions.

7.5 Amplification by gravitational lenses: microlensing and MACHOs

While gravitational lensing is commonly observed for massive lensing objects such as galaxies and clusters, distinct and separated images are not produced by individual stars, as the resolution of the best optical telescopes is simply not good enough, as shown in Example 7.1 below. While objects of typical stellar masses are too close to be resolved, distinguishing separate images is clearly possible for massive lensing objects, that is, for galaxies or galaxy clusters. Consider, for example, a cluster of mass $10^{14} M_{\mathrm{sun}}$, with $D_{\mathrm{LS}} = D_{\mathrm{L}} = D_{\mathrm{S}}/2 = 100\,\mathrm{Mpc}$. Then $\theta_{\mathrm{E}} \sim 65$ arc sec (assuming one can treat the cluster as a point mass), which is quite measurable.

Example 7.1 *Estimate the value of the Einstein radius for lensing by a stellar-type mass in the local galaxy, by considering the specific case of a pointlike lensing object of 10 solar masses, situated midway between the observer and a source at distance 2 pc (6 × 10^{16}m)—a typical interstellar distance. In this way discuss the possibility of observing separated images of individual stars with optical telescopes.*

Inserting the above numbers in (7.4), the value of the Einstein radius is found to be $\theta_{\mathrm{E}} = 0.32\mu\mathrm{rad} = 0.065$ arc sec. One has to compare this angle with the resolving limit of a telescope. An earth-bound optical telescope has a resolution of about 1 arc sec (5μ rad), and even that of the Hubble Space Telescope is good to only 0.1 arc sec. So, resolving separate optical images of sources lensed by objects of stellar mass is not feasible.

Even when the images of a source produced by gravitational lensing cannot be resolved, an amplification of the intensity may occur, in what is called a *microlensing event*. Suppose that a pointlike lensing mass is moving at velocity v normal to the line of sight, and that the source subtends an angle θ_{S} at the observer. All quantities are measured in the plane defined by the observer, lens, and source, with the notation of Fig. 7.3. In this case the angle θ_{S} will be a function of time, with a minimum value when the lens is at the closest distance of approach to the line of sight to the source. From Fig. 7.7, the right-angled triangle AS′L gives us $\mathrm{LS}'^2 = \mathrm{AS}'^2 + \mathrm{AL}^2$, where $\mathrm{LS}' = D_{\mathrm{L}}\theta_{\mathrm{s}}, \mathrm{AS}' = D_{\mathrm{L}}\theta_{\mathrm{s}}(\mathrm{min})$, and $\mathrm{AL} = vt$, where time t is measured from the moment of closest approach of the lensing object to the line of sight to the source. Dividing through by $(D_{\mathrm{L}}\theta_{\mathrm{E}})^2$ and with $x = \theta_{\mathrm{S}}/\theta_{\mathrm{E}}, x\,(\mathrm{min}) = \theta_{\mathrm{S}}\,(\mathrm{min})/\theta_{\mathrm{E}}$, we obtain

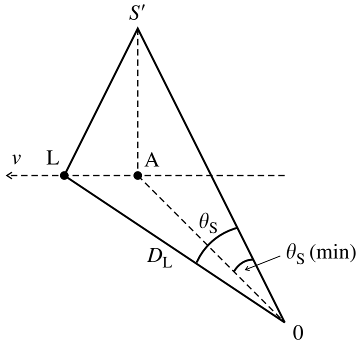


Fig. 7.7 A point lensing mass *L* moving with velocity *v* perpendicular to the line of sight. O is the observer and S′ is the projected position of the source in the plane of the lens.

$$\begin{aligned} x^2 &= x^2\,(\mathrm{min}) + \left(\frac{vt}{D_{\mathrm{L}}\theta_{\mathrm{E}}}\right)^2 \\ &= x^2\,(\mathrm{min}) + \frac{t^2}{T^2} \end{aligned} \qquad (7.6)$$

where in the second line we have defined $T = D_{\mathrm{L}}\theta_{\mathrm{E}}/v$. When the two images are not separated, there results an amplification of the (single) signal. From Liouville's theorem, the phase–space density, that is, the number of photons per unit solid angle, is unaffected by the imaging, so that if θ is the angle of the image, the amplification will be the ratio of solid angles, or

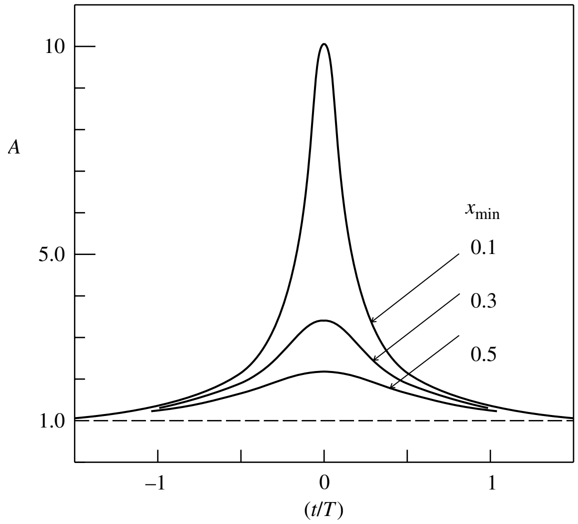


Fig. 7.8 Examples of the dependence of the amplification for microlensing events calculated from (7.8) for different values of $x(\text{min})$ as a function of t/T.

$A = \mathrm{d}\Omega/\mathrm{d}\Omega_S = \theta\,\mathrm{d}\theta/\theta_S\,\mathrm{d}\theta_S$. Since from (7.5)

$$\frac{\theta}{\theta_S} = \left(\frac{1}{2x}\right)\left[1 + \left(\frac{2}{x^2}\right) \pm x\sqrt{1 + \frac{4}{x^2}}\right] \tag{7.7}$$

it follows that, adding the amplitudes from the two (unresolved) images, the net amplification becomes

$$A = \frac{\left(1 + x^2/2\right)}{\left[x\sqrt{1 + x^2/4}\right]} \tag{7.8}$$

with x^2 defined in (7.6). Figure 7.8 shows how the signal depends on time, for a few cases of the ratio $x(\text{min})$. For $x(\text{min}) \ll 1$, the peak value of A is approximately equal to $1/x\,(\text{min})$. Figure 7.9 shows an example of a microlensing event, in which a massive dark object amplifies the light signal from a star in the Large Magellanic Cloud (a nearby mini-galaxy).

Example 7.2 *Calculate the typical time T for lensing by a pointlike object of mass* $0.1\ M_{\text{sun}}$*, moving at a velocity of* $v = 200\,km\,s^{-1}$ *normal to the line of sight, and situated half way to a source star at a distance of 50 kpc.*

Inserting these numbers in the above equations gives $\theta_E = \left[\left(4GM/c^2\right)(D_{LS}/D_S D_L)\right]^{1/2} = 6.2 \times 10^{-10}$ rdn and $T = D_L\theta_E/v = 2.39 \times 10^6\ \text{s} \sim 28$ days.

MACHOs is the name given to dark matter in the form of microlensing objects with masses of the order of stellar masses, in our galaxy. Typically their masses lie in the range 0.001–0.1 solar masses. Several hundred MACHOs have been observed, for example, by their microlensing of light from stars in the Large Magellanic Cloud, as in Fig. 7.9. A characteristic of these events is that the same amplification is observed in blue and red light, a fact which distinguishes them from variable stars. The reason for the achromaticity is clear. If the photon momentum is p, its effective gravitational mass is p/c, so that it will receive a

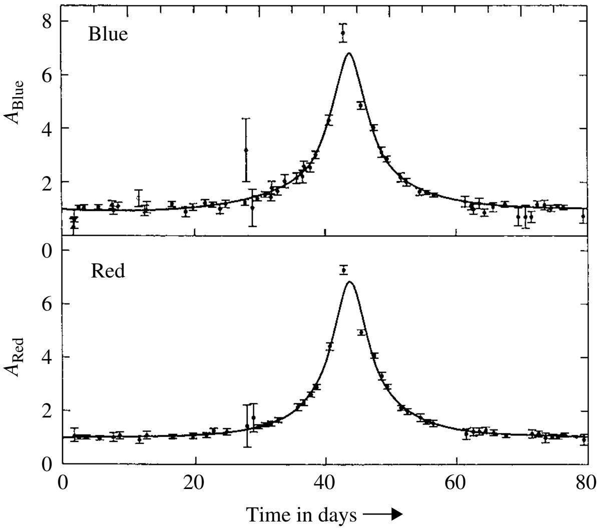


Fig. 7.9 Example of a microlensing event, the source being a star in the Large Magellanic Cloud, at a distance of some 50 kpc. Note the same signal is observed in blue and in red light. (After Alcock *et al.* 1993.)

transverse momentum $\Delta p \propto p$ from a gravitational field. Hence the deflection $\Delta p/p$ will be independent of wavelength h/p.

7.6 The lensing probability: optical depth

The probability that a particular source will undergo gravitational lensing as a measurable effect is called the *optical depth*. This is defined as the probability that at some instant of time, the line of sight to an individual star will be within the Einstein radius of a lens in the intervening distance. If N_{L} is the density of lenses per unit volume, and they are distributed uniformly, then since an Einstein ring extends over an area of $\pi(D_{\text{L}}\theta_{\text{E}})^2$ it follows that the optical depth will be

$$\tau = \int \pi D_{\text{L}}^2 N_{\text{L}}\, \mathrm{d}D_{\text{L}} \cdot \theta_{\text{E}}^2$$

where the integral extends from $D_{\text{L}} = 0$ to $D_{\text{L}} = D_{\text{S}}$. Substituting for θ_{E} from (7.4) and with $y = D_{\text{L}}/D_{\text{S}}$ and $\rho = N_{\text{L}}M$ for the mass density of lenses, the integral becomes

$$\tau = 4\pi \mathrm{G} \left(\frac{D_{\text{S}}}{c}\right)^2 \int \rho\left(y\right) \cdot y \cdot \left(1 - y\right) \mathrm{d}y$$

with y running from 0 to 1. If ρ is constant, this expression simplifies to

$$\tau = 2\pi \mathrm{G} \left(\frac{D_{\text{S}}}{c}\right)^2 \frac{\rho}{3} \tag{7.9}$$

which depends only on the distance to the source and the average mass density of lensing objects between the observer and the source. Inserting typical values of density for our galaxy, and considering sources near the periphery of the central bulge at $\sim 5\,\text{kpc}$, yields a value of $\tau \sim 10^{-7}$. Thus lensing will be a comparatively rare occurrence, and to detect dark matter in the form of 'dark stars'—that is, MACHOs as described above—one needs to examine the light curves of many millions of stars over months and years. This has required computerized search techniques, of the type first used in the old automated analysis systems in scanning bubble chamber film in particle physics experiments.

The magnitude of the magnification involved in microlensing varies inversely as the impact parameter between the lensed star and the MACHO, and so detection of microlensing typically involves examination of millions of stars, as discussed above. On the other hand, the Shapiro time delay falls off only logarithmically with impact parameter, so even with only a few thousand known pulsars, an effect may be detectable, although so far none has been claimed.

The present evidence is that MACHOs appear to account for only a small fraction of all the baryonic matter. The majority is in the form of stars, gas, and dust, of which by far the greatest contribution is from gas—commonly the very hot, X-ray emitting gas inhabiting galactic clusters, as discussed below. Certainly, MACHOs make only a trivial contribution to the energy density of dark matter. We now discuss some of the various candidates which have been proposed to constitute dark matter, as well as the experimental methods employed to search for them.

7.7 Baryonic dark matter

What is the nature of the dark matter which has been postulated to account for the phenomena described above? Some of the dark matter *must* be baryonic, since the value $\Omega_{\text{baryon}} \approx 0.04$ deduced from nucleosynthesis in the Big Bang is almost an order of magnitude larger than the closure parameter associated with visible stars, gas, and dust, of $\Omega_{\text{lum}} \sim 0.01$—see (5.31). In our own galaxy, some at least of this non-luminous baryonic dark matter has been accounted for in the form of compact halo objects (MACHOs) described above. However, X-ray studies of galaxy clusters reveal vast amounts of gas present between the galaxies in such clusters, and it seems possible that this accounts for almost half of the baryonic matter in the universe. Recent observations with the Chandra X-ray satellite, of absorption lines of oxygen and nitrogen, suggest that the other sources of the missing baryons are long filaments of gas associated with blazars (see Sections 9.10 and 9.14), which are AGN sources of TeV γ-rays.

There are no indications at present that more exotic baryonic objects such as mini black holes contribute significantly to the baryonic energy density. On the contrary, any primordial mini black holes of $M < 10^{11}$kg would have lifetimes less than the age of the universe, and in evaporating would emit Hawking radiation (as γ-rays)—see Sections 10.11 and 10.12. From the observed flux of γ-radiation from all sources one can set an upper limit on the energy density of such black holes of $\Omega_{\text{BH}} < 10^{-7}$.

In summary, baryonic matter makes only a small contribution to the overall density of the universe, and certainly less than 15% of the total estimated density of dark matter.

7.8 Neutrinos

The presently favoured hypothesis is that non-baryonic dark matter is composed of elementary particles, created at an early hot stage of the universe, and stable enough to have survived to the present day. The nature of such particles is presently a total mystery, although suggestions abound. To begin with, we can try to eliminate some known candidates. As indicated in Section 5.12, the neutrinos ν_e, ν_μ, and ν_τ and their antiparticles, together with electrons, positrons, and photons would have been produced prolifically in the early universe, and present in comparable numbers, according to the equilibrium reactions in (5.60):

$$\gamma \leftrightarrow e^+ + e^- \leftrightarrow \nu_i + \bar{\nu}_i$$

where $i = e, \mu, \tau$. As explained in Chapter 5, neutrinos would have frozen-out from further (weak) interactions with electrons or other matter when kT had fallen below about 3 MeV. Before better laboratory evidence on the smallness of neutrino masses became available, in the middle of the 1990s, the question arose as to whether neutrinos could be substantial—or even dominant—contributors to dark matter. In this respect they had the great advantage that they were at least known to exist. As discussed in Section 5.11, the relic microwave neutrino (plus antineutrino) number density would be comparable with the relic microwave photon number density, with a value (see (5.67)):

$$N_\nu = \left(\frac{3}{11}\right) N_\gamma = 113\,\mathrm{cm}^{-3} \tag{7.10}$$

per neutrino flavour, to be compared with a number density of 411 cm^{-3} for the microwave photons. We note from this number density that the total energy density of neutrinos would be equal to the critical density (5.26) if the sum of the masses of the three flavours had the value

$$\sum_{e,\mu,\tau} m_\nu c^2 = 47\,\mathrm{eV} \tag{7.11}$$

So neutrinos with masses in the few eV mass range could make significant contributions to dark matter. However, evidence from neutrino oscillations (Section 4.2) indicates very much smaller masses than in (7.11), below 0.1 eV/c^2 as judged from the mass differences. Another problem with neutrinos as dark matter candidates is that they would constitute 'hot' dark matter. With the critical temperature of $kT \sim 3$ MeV, neutrinos were relativistic when they decoupled from other matter and also when the structures in the universe were forming. Consequently, they would stream away rapidly and, just like the photons, tend to iron out any primordial density fluctuations. So if large-scale structures are to form, early computer simulations indicated that the fraction of dark matter which is 'hot' could only be of order 30% or less. All this will be discussed in detail in Sections 8.9–8.11.

Aside from the question of relic neutrinos forming dark matter, the existence inferred above of some 340 neutrinos in each cubic centimetre throughout space, with energies in the milli-electron volt range, poses a truly formidable challenge to the experimentalist to detect them. After microwave photons, the relic neutrinos are, as far as we know, by far the most prolific particles in the universe, but there seems to be no obvious way to show their existence. Since the value $T_\nu = 1.9$ K given in Section 5.11 implies, together with the data on mass differences in Section 4.2, that the relic neutrinos (or at least two of the three flavours) are non-relativistic, their de Broglie wavelengths will be of order 0.5 mm, so they could be coherently scattered by sizeable lumps of matter. However, so far no really plausible method of detecting them has been found, although several ideas have been put forward. For example, the Pauli principle could cause suppression of weak decays resulting in emitted neutrinos trying to occupy the same cells of phase space as the relic neutrinos. It is also remotely possible that, if ultra-high energy (10^{23} eV!) neutrino sources actually exist, their spectrum might show a detectable dip where such neutrinos interact with the relic (anti)neutrinos to form the resonance $\nu + \bar{\nu} \rightarrow Z^0$. But all such ideas seem to be for the far future.

The foregoing discussion of course applies to the 'light' neutrinos ν_e, ν_μ, and ν_τ we are familiar with in the laboratory. There is the additional possibility of extremely massive (GUT scale) Majorana neutrinos being created at a very early stage of the universe, as discussed in Section 4.4. These could have played a vital role in the development of the universal matter–antimatter asymmetry. However, such massive neutrinos would be unstable and have disappeared by decay, and could not have contributed to dark matter today.

7.9 Axions

The axion is a very light pseudoscalar particle (spin-parity 0^-) originally postulated in connection with the absence of CP violation in strong interactions (quantum chromodynamics,QCD). In principle, complex phases can occur in the quark wavefunctions in QCD, and these would be T violating or CP violating (as indeed they are in the weak interaction sector). However, the upper limit to the electric dipole moment of the neutron is nine orders of magnitude less than strong CP violation predicts. To cancel this undesirable feature and to account for the smallness of any possible violation, Peccei and Quinn (1977) proposed a new global U(1) symmetry, spontaneously broken at some very-high-energy scale, and giving rise to an associated boson (a so-called Goldstone boson) called the axion, which receives a small mass via non-perturbative (instanton) effects at phase transitions on the QCD scale (200 GeV). The axion, just like the pseudoscalar neutral pion, can decay into two photons, at a rate determined by the extremely weak coupling $1/f_a$ to other particles, where f_a is the Peccei-Quinn energy scale (all quantities here are expressed in 'natural units' $\hbar = c = 1$). The axion mass is given by the formula

$$m_a \approx 0.5 \frac{m_\pi f_\pi}{f_a} \approx \frac{6\,\text{eV}}{\left[f_a/\left(10^6\,\text{GeV}\right)\right]} \tag{7.12}$$

where $f_\pi = 93$ MeV is the pion decay constant. Thus a value of $m_a = 1$ (0.01) eV/c^2 would correspond to $f_a = 6 \times 10^6$ (6×10^8) GeV. The actual lifetime for axion decay to two photons is proportional to $1/m_a^5$ and exceeds the age of the universe for any mass below 10 eV/c^2. If they exist, axions would therefore, like the primordial photons and neutrinos, survive as relics of the Big Bang.

The earliest limits on axions were set from astrophysics. Owing to their two-photon coupling, axions could be produced in stars by their conversion from photons via the Primakoff effect, in which a photon interacts with the Coulomb field of a nucleus, $\gamma + \gamma \to a$. Owing to their extremely weak coupling, axions would be freely emitted from and greatly enhance the cooling rate of red giant stars in globular clusters to an unacceptable level, since the energy loss would have to be compensated by increased nuclear fusion and a shortening of the lifetime of the star, and a reduction in the numbers visible at any one time. It would certainly be inconsistent with conventional stellar models, which are able to give quite successful descriptions of stellar evolution (see Chapter 10). These considerations gave an upper limit to the coupling $1/f_a$, or equivalently to the axion mass, of $m_a < 0.01$ eV/c^2.

The decay of the axion to two photons implies that in a suitable magnetic field (supplying an incoming photon), one should be able to observe the conversion of an axion to a photon. The CAST (CERN Axion Solar Telescope) experiment (Zioutas *et al.* 2005) sought to detect axions emitted from the Sun, again produced there by the Primakoff effect of converting a photon to an axion in the Coulomb field of a nucleus in the Sun. Restricting observations to the core of the Sun, where photons have keV energies, this experiment looked for the X-rays which would result from the conversion of solar axions back into photons, using a $9T$ superconducting magnet. The absence of any signal gave a limit on the axion mass similar to that above, of $m_a < 0.02$ eV/c^2.

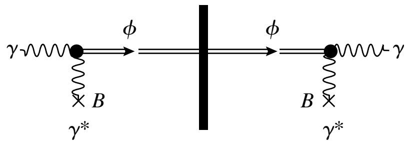


Fig. 7.10 'Photons through a wall' experiment. On the left, an incoming photon interacts with a photon of a very strong magnetic field, converting to a very weakly interacting axion ϕ, which passes through a wall. On the right-hand side, the axion converts back into a photon in a second magnetic field.

Attempts have been made at direct observation of axions created in laboratory experiments, using the 'photons through a wall' method (see Fig. 7.10). A (plane-polarized) laser beam traverses a magnetic dipole field, converting photons to axions, which then easily pass through a wall, on the other side of which the axions are converted back to photons in a second dipole field (Cameron *et al.* 1993, Ehret *et al.* 2007). Future experiments of this type will certainly be crucial in the axion search.

The extreme smallness of the axion mass deduced from the above limits would appear to preclude it as a serious dark matter candidate, if one uses the same arguments that led to the mass limit for neutrinos quoted above. However, the very weakness of the axion coupling means that the axions formed in the early inflationary stage of the universe never get into thermal equilibrium with other particles, and the 'freeze-out' arguments applied to neutrinos are not relevant. Instead, axions would have formed as a boson condensate of cold dark matter.

In order to account for dark matter, that is, to reach an energy density of order the critical density, one requires axion masses of at least $10^{-5} - 10^{-3}$ eV.

7.10 Axion-like particles

The above remarks apply to conventional axions, as first envisaged in the 1970s. As the search for other types of dark matter particles (WIMPs as described below) has failed to find any evidence, the axion hypothesis, and variations on it, has found more favour. Extensions of the Standard Model, grand unification theories and supergravity theories all have room to accommodate axion-type particles. Indeed, the definition of axion can encompass particles with somewhat different properties, for example, scalar rather than pseudoscalar particles, and particles which do not couple directly to quarks or leptons. Furthermore, depending on the properties assumed, such particles can form either cold or hot dark matter. They are referred to as axion-like particles (ALPs) and the couplings are no longer related to the axion masses as in (7.12).

In the so-called hadronic axion models, axions have no first order coupling to leptons or quarks, in which case the foregoing mass limits do not apply. The axion is coupled to hadrons, for example, to pions in the process $a + \pi \rightarrow \pi + \pi$. In this case, thermalization of the axions after the temperature falls below the QCD scale of $\sim 100\,\text{GeV}$ can take place, and similar arguments to those for neutrinos on freeze-out rates will then apply. The present axion intensity would be of the same order as for relic neutrinos, and if the axion mass is in the eV mass range, the axion could again make an important contribution to dark matter. However, the crucial experiments definitely demonstrating the existence of axions, hot or cold, still remain to be done.

7.11 Weakly interacting massive particles

The other favoured hypothesis for dark matter particles is that they are weakly interacting massive particles (WIMPs), moving with non-relativistic velocities at the time of freeze-out and thus constituting cold dark matter.

First, however, we should ask if these particles could be massive neutrinos. If they were, and had masses such that they were still relativistic at freeze-out, the closure parameter would increase with neutrino mass m_ν as in (7.11), and for $m_\nu \sim 1$ MeV, would reach the unphysical value of $\Omega \sim 10^4$. For higher masses, the heavy neutrinos would become non-relativistic at decoupling, and as shown below, this then results in $\Omega \propto 1/m_\nu^2$ instead of $\Omega \propto m_\nu$. The closure parameter falls back below $\Omega = 1$ for a mass above 3 GeV. So the fact that the measured value of Ω is not large compared with unity, excludes the mass range 50 eV–3 GeV, while experiments at the LEP electron–positron collider at CERN prove conclusively that there are no 'extra' conventional neutrinos of mass below $(1/2)M_Z = 45\,\text{GeV}$. Conventional neutrinos of $m_\nu > 45\,\text{GeV}$, if they existed, would, however, make no significant contribution to Ω (< 0.01, from the $1/m_\nu^2$ dependence). The dependence of the closure parameter on mass is illustrated in Fig. 7.11, drawn for WIMPs generally, that is, for massive non-relativistic particles with conventional weak couplings.

For several reasons, it is considered that *supersymmetric (SUSY) particles* could be the most likely WIMP candidates. As described in Section 4.5, such SUSY particles are expected to be created in pairs, with opposite values $R = \pm 1$ of a conserved quantum number called R-parity. Heavier SUSY particles would decay to lighter ones in R-conserving processes, ultimately ending up with the lightest supersymmetric particle (LSP), which we denote by the symbol χ. This particle is assumed to be stable and therefore to have survived from the primordial era of the universe. The LSP is usually identified with the *neutralino,* a neutral fermion which is the lightest of the states arising from a linear combination of the photino, zino, and two higgsinos (see Table 4.1). Since such anomalously heavy particles are not constituents of atoms or nuclei, they cannot have either electromagnetic or strong coupling and are assumed to interact only weakly. Although neutralinos are stable, they can of course disappear by annihilation with their antiparticles, which will have been generated with the same abundance as the particles. There are many free parameters in SUSY models, which means that the LSP mass as well as the annihilation cross-section and cosmological abundance can vary over quite wide ranges, and it is probably this flexibility which is part of the attraction of such models.

Let us now look more closely at the constraints which WIMP models have to fulfil. First, we are searching for *cold* dark matter, since this must form the bulk of all dark matter if one is to successfully account for the development of structures in the universe, as explained in Chapter 8. So the WIMPs must be non-relativistic when they 'freeze-out'. This freeze-out occurs when the rate of $\chi\bar{\chi}$ annihilation falls below the expansion rate, that is, when

$$N \langle \sigma v \rangle \leq H \tag{7.13}$$

where N is the WIMP number density, v is the relative velocity of particle and antiparticle, σ is the WIMP–antiWIMP annihilation cross-section, and H is the Hubble parameter at the time of freeze-out. It will be seen that the WIMP abundance varies inversely as the annihilation cross-section, so that weaker interactions lead to earlier freeze-out and consequently higher densities and larger contributions to the closure parameter. Since the WIMPs are massive and non-relativistic, $M \gg T$ where M is the WIMP mass and T is the temperature at freeze-out in energy units. Then the density will be given by the Boltzmann relation (see Problem 5.3):

$$N(T) = \left(\frac{MT}{2\pi}\right)^{3/2} \exp\left(\frac{-M}{T}\right) \tag{7.14}$$

The exact value of the $\chi\bar{\chi}$ annihilation cross-section is of course unknown, but if it is of the same order as the weak cross-section then on dimensional grounds we could set $< \sigma v > \sim G_F^2 M^2$—see (1.27). Inserting (7.14) in (7.13) and with $H = 1.66g *^{1/2}\, T^2/M_{\mathrm{PL}}$ as given by (5.59) in the radiation-dominated universe, the freeze-out condition becomes

$$(MT)^{3/2} \exp\left(\frac{-M}{T}\right) G_{\mathrm{F}}^2 M^2 \leq \frac{fT^2}{M_{\mathrm{PL}}} \tag{7.15}$$

where f includes the numerical constants involved and is of order 100. The Fermi constant squared is $G_F^2 \approx 10^{-10}$ GeV^{-4} while the Planck mass $M_{PL} \approx 10^{19}$GeV. Inserting these numbers, one can easily solve numerically for the value of $P = M/T$ at freeze-out. It varies slowly and logarithmically, from around $P = 20$ for $M = 1$ GeV to around $P = 30$ for $M = 100$ GeV. In the following we take this ratio as a constant, $P = 25$. Now going back to (7.13), and recalling that the expansion parameter $R \propto 1/T$, the WIMP density $N(0)$ today, when the CMB temperature is $T_0 = 2.73$ K, will be

$$N(0) \sim \frac{(T_0/T)^3 \times \left(T^2/M_{\mathrm{PL}}\right)}{\langle \sigma v \rangle} \tag{7.16}$$

The corresponding WIMP energy density will be

$$\rho_{\mathrm{WIMP}} = MN(0) \sim \frac{PT_0^3}{(M_{\mathrm{PL}} \langle \sigma v \rangle)} \sim \frac{6 \times 10^{-31}}{\langle \sigma v \rangle} \,\mathrm{GeV\,s^{-1}}$$

with σv in cm^3s^{-1}. Dividing by the critical energy density $\rho_c = 3H_0^2c^2/8\pi\text{G} \approx 5 \times 10^{-6}\,\text{GeV}\,\text{cm}^{-3}$ from (5.26) we get for the closure parameter

$$\Omega_{\text{WIMP}} = \frac{\rho_{\text{WIMP}}}{\rho_c} \sim \frac{10^{-25}\,\text{cm}^3\,\text{s}^{-1}}{\langle\sigma v\rangle} \tag{7.17}$$

At freeze-out the WIMP velocity will be given by $Mv^2/2 = 3T/2$, or $v/c \sim (3/P)^{1/2} \sim 0.3$ so that (7.17) indicates that an annihilation cross-section of order $10^{-35}\,\text{cm}^2$ would lead to a closure parameter of order unity. It is perhaps quite remarkable that this cross-section is of the order of magnitude expected for the weak interactions, since there is no *a priori* connection between the closure density of the universe and the Fermi constant. In any case, if this is a mere coincidence it is a bonus, in the sense that one does not have to invent new couplings as well as new particles in trying to account for dark matter of this type.

However, as indicated previously for neutrinos, for the conventional weak coupling the annihilation cross-section will rise as M^2 and therefore the closure parameter falls as $1/M^2$, and at high values of M the WIMPs could make no substantial contribution to the energy density. This state of affairs holds until the WIMP mass becomes comparable with or larger than the mass of the mediating weak bosons, W and Z. Then, as indicated in (1.9) the boson mass in the propagator term becomes less important. In the electroweak model, $g_W^2 = G_\text{F}M_W^2 \sim \alpha = 1/137$, and the $\chi\bar{\chi}$ annihilation cross-section, when $M \gg M_W$, will be of order $g_W^4/M^2 = \alpha^2/M^2$. It falls off as $1/s$ where $s \sim 4M^2$ is the square of the centre-of-momentum system (CMS) energy, just as for the electromagnetic cross-section in Fig. 1.9. Consequently, the value of Ω_{WIMP} now *increases* as M^2 instead of decreasing as $1/M^2$, as shown in Fig. 7.11. So, WIMP masses even in the TeV mass range could be important dark matter candidates, and the flexibility in the couplings in the various SUSY models means that a wide range of WIMP masses is possible.

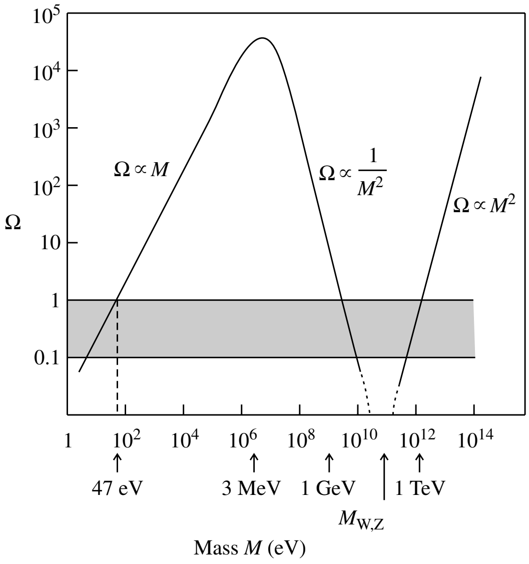


Fig. 7.11 Variation of the closure parameter with WIMP mass, assuming conventional weak coupling. The shaded region, corresponding to $\Omega = 0.1 - 1$, is that in which the contribution to the closure parameter from massive neutrinos or WIMPs must lie, thus excluding the range of masses 100 eV–3 GeV. Accelerator experiments suggest that WIMPs must have masses exceeding $M_Z/2 = 45\,\text{GeV}$, otherwise Z bosons could decay into WIMP–antiWIMP pairs. However, for masses which are large compared with the Z boson mass, the weak cross-section falls rapidly because of propagator effects, so that WIMPs in the TeV mass range are possible dark matter candidates, depending on the precise values of the WIMP coupling.

7.12 Expected WIMP cross-sections and event rates

There are two distinct possibilities for detection of WIMPs. Direct detection of dark matter relies on observation of the scattering or other interaction of the WIMPs inside the detector, while indirect detection relies on the observation of the annihilation products of WIMPs in, for example, the halo of the galaxy, or as a result of their accumulation in the core of the Sun or the Earth. In the latter cases of course, the only secondary products which could be detected would be neutrinos. In fact, no evidence for an extra flux of high-energy neutrinos from the direction of the Sun or from the Earth's core has ever been found.

In the case of direct detection, the WIMP rate may be expected to exhibit some angular and time dependence. For example, if WIMPs predominantly populate the galactic halo, there might be a daily modulation because of the shadowing effects of the Earth when turned away from the galactic centre. An annual modulation in the event rate would also be expected as the Earth's orbital velocity around the Sun adds to or subtracts from the velocity of the Solar System with respect to the galactic centre, so that both the velocity distribution of WIMPs and the cross-section for detection change with time.

We now discuss the expectations for elastic scattering of WIMPs by nuclei in the detector, signalled by the nuclear recoil. The WIMP velocities are expected to be of the order of galactic escape velocities, that is, $v \sim 10^{-3}c$, so that we can use non-relativistic kinematics. Then if $E = Mv^2/2$ is the kinetic energy of a WIMP of mass M, colliding with a nucleus of mass $M_N = mA$ where A is the mass number and m the nucleon mass, it is straightforward to show that the total CMS energy is (if required, refer to Chapter 2 for relativistic transformations):

$$\begin{aligned}\varepsilon &= \left[(M + M_{\rm N})^2 + 2M_{\rm N}E\right]^{1/2}\\ &\approx [M + M_{\rm N}]\left[1 + \frac{M_{\rm N}E}{(M_{\rm N} + M)^2}\right]\end{aligned} \tag{7.18}$$

where in the second line we use the fact that $E \ll M_N$ or M. If p^* denotes the (equal and opposite) momentum of each particle in the CMS, then in the non-relativistic approximation

$$\varepsilon = \left(M_{\rm N} + \frac{p^{*2}}{2M_{\rm N}}\right) + \left(M + \frac{p^{*2}}{2M}\right) \tag{7.19}$$

so these two equations give

$$p^{*2} = \frac{2\mu^2 E}{M} = \mu^2 v^2 \tag{7.20}$$

where $\mu = M_{\rm N}M/(M_{\rm N} + M)$ is the reduced mass. The laboratory kinetic energy E_r of the recoiling nucleus is maximum when its CMS vector momentum is reversed in the collision, so that it is scattered in the forward direction with laboratory momentum $2p^*$ and E_r (max) $= 2p{*}^2/M_{\rm N} = 2\mu^2 v^2/M_{\rm N}$. This has a value varying from $v^2 M_{\rm N}/2$ when $M_{\rm N} = M$ to $2v^2 M_{\rm N}$ when $M \gg M_{\rm N}$. Since the CMS angular distribution at these low velocities will be isotropic, the recoil energy distribution will vary uniformly between zero and E_r(max). So with

$v \sim 10^{-3}c$ and $M_{\mathrm{N}} \sim A$ GeV, we obtain recoil energies $E_r \sim A$ keV or less. Hence, a sensitive detector is needed to observe such small recoil energies.

The scattering cross-section of the target nucleus depends on details of the SUSY model parameterization. For guidance we again assume a conventional weak cross-section. From (1.18) with $|T_{\mathrm{if}}| = G_{\mathrm{F}}$ the cross-section per target nucleus will be

$$\sigma \approx \frac{G_{\mathrm{F}}^2 p^{*2} K}{\pi v_r^2} = \frac{G_{\mathrm{F}}^2 \mu^2 K}{\pi} \tag{7.21}$$

where the relative velocity of incident particle and target nucleus in the CMS is $v_r = v = p^*/\mu$. The quantity K is a numerical model-dependent factor. For spin-independent coupling, the scattering amplitudes from the different nucleons in the target nucleus should add coherently, so that K will contain a factor A^2. However, the momentum transfer is of order $p^* = \mu v \sim 10^{-3}A$ GeV, while the nuclear radius $R = 1.4A^{1/3}$ fm $\sim 7A^{1/3}$ GeV^{-1}. The nucleus can only recoil coherently if $p^*R \ll 1$, or $A \ll 50$, otherwise K will contain a suppression factor (the square of the so-called form-factor).

The other possibility is spin-dependent (axial–vector) coupling, for which the amplitudes from different nucleons do not add since most of the nucleon spins cancel out, and the cross-section is smaller by a factor of order A^2 than that for coherent scattering. As examples, for WIMPs identified with sneutrinos (see Table 4.1) the interaction is scalar and coherent, while if the WIMP is the LSP (neutralino) with spin $1/2$, the interaction will be mostly incoherent.

The event rate to be expected depends on the WIMP number density and the scattering cross-section. Because of their gravitational concentration in the galaxy and particularly the disc and halo, the WIMP energy density in the solar system is estimated to be some 10^5 times that in the universe at large, at $\rho_{\mathrm{WIMP}} \sim 0.3$ GeV cm^{-3}, yielding a flux of $\varphi_{\mathrm{WIMP}} \sim 0.3v/M$ cm^{-2} s^{-1}, where the WIMP mass M is in GeV. The reaction rate per target nucleus will be $W = \sigma\varphi_{\mathrm{WIMP}}$ as in (1.14) and the event rate per unit target mass from (7.21) will be

$$R = \frac{W}{M_{\mathrm{N}}} \sim \frac{10K}{AM} \text{ events kg}^{-1}\text{ day}^{-1} \tag{7.22}$$

Typical values of $M = 100$ GeV and $A = 20$ predict $R \sim 0.01$ events per kg day for incoherent scattering and $R \sim 1$ per kg day for coherent. As indicated below, present upper limits are well below these figures. The cross-sections and rates of course depend on the many free parameters in SUSY models, if we assume that WIMPs are supersymmetric particles, and so the above numbers are indicative only: but they suffice to emphasize the severe experimental problems of detecting signals from low-energy recoils at extremely low rates, against cosmic ray and radioactive background effects.

7.13 Experimental WIMP searches

Direct detection of WIMPs via the recoil of the scattering nucleus has been attempted by a number of different methods. The ionization from the recoil as it traverses the detector material can be recorded as a pulse in a semiconductor counter (Ge or Si), which have excellent sensitivity to recoils in the keV energy range, or in the form of scintillation light from scintillating materials such as NaI or liquid Xe. However, the bulk of the energy lost by the recoil will appear in the form of lattice vibrations (phonons) in the medium. These can be recorded through cryogenic detectors operating at low temperature (< 1 K). The phonon pulse results in a local rise in temperature, which will affect the resistance of a thermistor attached to the detector and can be recorded as a voltage pulse. The phonon pulses are very slow in comparison with electrical pulses from ionization, and therefore random background noise can be more of a problem.

As stated before, the signals from WIMPs have to be distinguished from those due to background radioactivity and the interactions of cosmic ray induced neutrons and photons. For this reason, emphasis has to be on very pure materials and on locating the detectors deep underground to reduce the cosmic ray muon flux. The separation of genuine from background events can be achieved in several ways. For example, the energy spectrum and event rate of recoils will be different for detectors with nuclei having different A values and/or different nuclear spins. Some discrimination is also possible on the basis of pulse length in scintillators. Electrons produced from photon or radioactive background have longer pulse lengths than nuclear recoils of the same energy. Similarly, the ratio of ionization energy loss to lattice (phonon) energy loss is also different for recoil nuclei and for electrons. Finally, WIMP recoils should show a small seasonal dependence of the signal. The latter arises from the fact that the Sun orbits the Galaxy with $v \sim 200$ km s^{-1}, while the Earth orbits the Sun with $v \sim 30$ km s^{-1}. The two velocities add vectorially to give a maximum in summer (on 3 June) and minimum in winter. There results a small annual change in WIMP fluxes, detector cross-sections, and event rates, of the order of 5%.

Figure 7.12(a) shows early (2002) experimental upper limits on WIMP-nucleon scattering cross-sections, assuming coherent nuclear scattering, taken from the first edition of this text. For small WIMP masses, the limit at first decreases with increasing WIMP mass, because more of the recoil energies are above the detection threshold; after passing through a minimum, the cross-section limit then increases again, as the flux of WIMPs for a given cold dark matter closure parameter decreases with increasing WIMP mass.

These and later limits already exclude the cross-section ranges expected for LSPs in some versions of the supersymmetric models. So far, only one experiment has claimed a signal in the form of an annual modulation. Using a large (100 kg) NaI detector, the DAMA group reported a 5% annual modulation for low recoil energies, less than 6 keV, with a significance level of about 2 standard deviations. However, the EDELWEISS experiment with a cryogenic Ge detector, and the ZEPLIN experiment using liquid Xe, set limits which were apparently incompatible with the DAMA result. More recent results place the upper cross-section limits at least an order of magnitude lower than those shown in Fig. 7.12(a). For example, the CDMS experiments (Fig. 7.12(b)) with cryogenic germanium and silicon detectors find an upper cross-section limit of 2.5×10^{-7} pb (for a WIMP mass of order 60 GeV/c^2). These limits assume coherent nuclear scattering. If it is spin-dependent, the disagreement with the DAMA results (now with a 250kg detector and an 8σ effect) is less clear. The present limits only exclude a part of the parameter range of supersymmetric

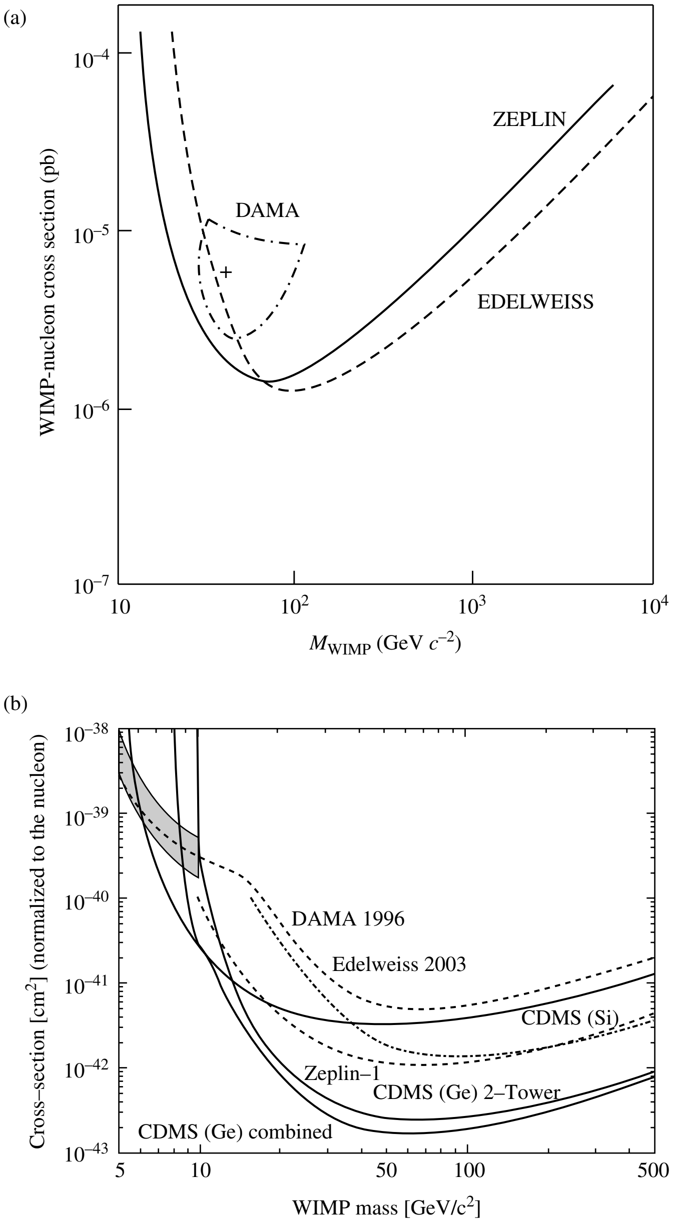


Fig. 7.12 (a) Upper limits on the WIMP-nucleon scattering cross-section as a function of WIMP mass from the EDELWEISS (Benoit *et al.* 2002) and ZEPLIN (Smith 2002) experiments. The cross-section inferred by the DAMA group (Bernabei *et al.* 2002) from annual modulation is shown by the closed contour (see also Bernabei *et al.* 2008). (b) Cross-section limits from the more recent CDMS experiment (Akerib *et al.* 2006) are about one order of magnitude smaller.

models, and the search needs to be continued with detectors of ever greater mass and sensitivity.

7.14 Dark energy: high redshift supernovae and the Hubble plot at large z

As stated in Chapter 5, the total energy density $\rho_{\rm tot}$ appearing in the Friedmann equation (5.11) may have three separate sources—matter, radiation, and vacuum energy density—as in Table 5.2. For non-relativistic matter, $\rho \propto R^{-3}$ while for radiation or any ultra-relativistic particles, $\rho \propto R^{-4}$. In either case, as indicated from this table, the density falls off with time as $1/t^2$. On the other hand, the vacuum energy density—if indeed that is the source of dark energy—is constant, so that however small it may be relative to other forms of energy density at early times, eventually it must begin to dominate at large enough values of t. From the expression (5.46b) for the deceleration parameter q it is apparent that if at some epoch the vacuum energy density $\rho_\Lambda > \rho_r + \rho_m/2$, the universe will *accelerate.*

7.14.1 Type Ia supernovae

The evidence for a substantial dark energy component comes from several sources: galaxy redshift surveys; gravitational lensing, the age of the universe (see Example 5.3) and particularly the age estimates of globular clusters described in Chapter 10; but most dramatically and originally, from the measurement of the Hubble flow at large redshifts, from the analysis of Type Ia supernova luminosities. In 1997 two independent investigations made the startling discovery that, although in the distant past, the Hubble expansion was decelerating because of the braking effect of the gravitational interaction of matter, this has more recently been replaced by an acceleration (Riess *et al.* 1998, Perlmutter *et al.* 1999). The data for this conclusion come from one of several different types of exploding stars called supernovae, all of which have the common feature that they become unstable and explode when the mass of the stellar core exceeds the Chandrasekhar limit and implodes, that is, when gravitational pressure exceeds electron degeneracy pressure (see Chapter 10). Type II supernovae, and also Types Ib and Ic, are associated with the final stages of thermonuclear fusion and gravitational collapse of massive stars, once the core exceeds the Chandrasekhar limit, and their subsequent transformation to neutron stars and black holes.

Type Ia supernovae, which are what concern us here, are distinguished from the other types by the presence of lines from silicon and absence of lines from hydrogen in their spectra. The mechanism involved in this case is also different. It is believed that they develop from stars which have burned all their hydrogen, and have reached the white dwarf stage with a carbon/oxygen core, which, however, is not massive enough to provide the high temperatures needed to permit the thermonuclear fusion of still heavier elements. The flash is due to the explosion of the white dwarf, which is part of a binary system. It has steadily accreted matter from its main sequence companion, until the core eventually exceeds the critical Chandrasekhar mass and implodes down to nuclear density, with a huge release of gravitational energy. The result is that in a matter of seconds the stellar material is converted largely to heavier elements such as silicon, nickel, and iron by rapid thermonuclear fusion, with a tremendous release of nuclear binding energy and the subsequent explosion. The dispersed nickel nuclei subsequently decay to cobalt and iron over a period of months, setting the timescale for the (roughly) exponential decay of the light curve (see Fig. 10.8 for an example from a Type II supernova).

The light output from a Type Ia supernova typically grows over a period of a few weeks, before reaching a maximum and thereafter falling off exponentially. There are some variations in the maximum light output between different supernovae, and the peak luminosity depends on the timescale τ to reach the maximum (varying as $\sim \tau^{1.7}$). The brighter supernovae originate from more massive stars and the ensuing fireball has to expand for longer in order for the opacity to drop enough to allow the photons to escape. After making this empirical correction based on the 'width' of the light curve, the estimated total light output from different supernovae shows remarkably small dispersion, of the order of 10% only.

In Type II supernovae, discussed in Chapter 10, the vast bulk (99%) of the energy release is in the form of neutrinos, but this is not expected to be the case for Type Ia. Thus, although they originate from smaller stars, a greater fraction of the output is in the form of light, so their (photon) luminosities are comparable with those of Type II.

7.14.2 The Hubble plot at large redshifts

Before describing the experimental results, let us first ask how different cosmological parameters can change the slope of the Hubble plot as a function of redshift. The actual plot is made of the distance as estimated from the luminosity or apparent magnitude of the star, that is, the so-called luminosity distance D_L defined in (5.5). The expected value of D_L can be calculated as a function of redshift z from the presently measured value of the Hubble parameter H_0, and assumed values of the various contributions to the closure parameter Ω, as defined in Section 5.5. First we recall that the true coordinate distance $D(z) = R(0)r$ to an object at redshift z and co-moving distance r is given by equation (5.44b). The luminosity distance $D_\mathrm{L}(z)$ in (5.5) and (5.6) is given by the luminosity L in terms of the power P radiated isotropically by the source:

$$L = \frac{P}{4\pi\,[R\,(0)\,r]^2} \times \frac{1}{(1+z)^2} = \frac{P}{4\pi D_\mathrm{L}^2} \tag{7.23a}$$

so that

$$D_\mathrm{L} = (1+z)\,D\,(z) \tag{7.23b}$$

In (7.23a), one factor of $1/(1+z)$ arises because pulses of light emitted from the source at redshift z, over a time interval Δt, will arrive at the detector over a stretched time interval $\Delta t(1+z)$. The second $1/(1+z)$ factor arises because the energy per photon at emission has been red-shifted downwards by the time it reaches the detector. Equations (7.23) and (5.44) give expressions for the luminosity distance in terms of the Hubble parameter H_0 and the contributions to the energy density from matter, radiation, vacuum/dark energy, and curvature terms. Since, in dealing with the supernova results, we are concerned with redshifts of order unity or less, we can certainly neglect radiation, since it is important only at very large redshifts ($z > 1000$). It may also be remarked here that at high redshift, allowance must also be made for the fact that the supernova decay curves themselves will be 'stretched' by the time dilation factor $(1+z)$.

The expected results for different scenarios follow from straightforward integration of (5.42). Table 7.1 gives the resulting expressions for the dimensionless quantity $D_\mathrm{L}\,(z)\,[H_0/c]$.

Figure 5.3 gave the results from the measurements of supernovae at low redshifts ($z < 0.1$), using Cepheid variables in the same galaxies to calibrate

Table 7.1 Luminosity distance versus redshift

Dominant component	Ω_m	Ω_Λ	Ω_k	$D_L H_0/c$
Matter (Einstein–de Sitter universe)	1	0	0	$2(1+z)[1-(1+z)^{-1/2}]$
Empty universe	0	0	1	$z\,(1+z/2)$
Vacuum	0	1	0	$z(1+z)$
Flat, matter + vacuum	0.24	0.76	0	Numerical integration giving best fit to data (see Fig. 7.14)

the distance/luminosity scale. The results for this region of z indicated a very uniform Hubble flow with $H_0 = 72$ km s^{-1} Mpc^{-1}. The fact that the data fall on a straight line with little dispersion gives confidence that the normalization methods employed are adequate.

Because of this reproducibility and their extreme brightness, which allows one to probe to large distances and redshifts, Type Ia supernovae have come to be regarded as ‘standard candles’, so that their brightness or apparent magnitude, when coupled with the decay curve, fixes the integrated luminosity and thus the distance from the Earth. However, such events only occur at the rate of order one per century per galaxy. The method employed in the High z Supernova Search, for example, was to scan a strip of the sky containing around ten thousand galaxies, using the Hubble Space Telescope, then to repeat the survey 3 weeks later and by taking the difference, to detect the dozen or so supernovae which had developed in the meantime. Once identified, their light curves could be studied in detail. The early pioneering experiments in this field were those of the High z SN Search Team (Riess *et al.* 2000) and the Supernova Cosmology Project (Perlmutter *et al.* 1999). Since then, various experimental groups have contributed data, using both the Hubble Space Telescope and ground-based telescopes.

Figure 7.13 shows typical results by Clochiatti *et al.* (2006), where data from several experiments have been included. The top panel shows values of the distance modulus or logarithm of the luminosity distance in (5.6), plotted against the logarithm of the redshift z. The predicted variation for different cosmological parameters is shown by the curves, which however, are too close together to interpret easily. Notice that, even for a non-accelerating universe, where H is obviously constant, the plot in Fig. 7.13 is *not* a straight line, because of the way in which distance (or magnitude) is measured from the observed brightness of the source in an expanding universe. The lower panel shows the *difference* in magnitude, as compared with that in an empty, non-accelerating universe. An empty universe in this plot then corresponds to a horizontal line at zero on the vertical scale, an accelerating universe will show a curve with an upward slope and a decelerating universe, one with a downward slope. The points indicate a change from an accelerating universe at small z values to a decelerating one for $z > 0.5$. The reader can easily check from a rough numerical integration of (5.39) that this was about 5 billion years ago, that is, when the universe was two-thirds of its present age.

Figure 7.14 again shows the results on the differential magnitude (comparing with an empty universe) from Riess *et al.* (2004), where the numerous

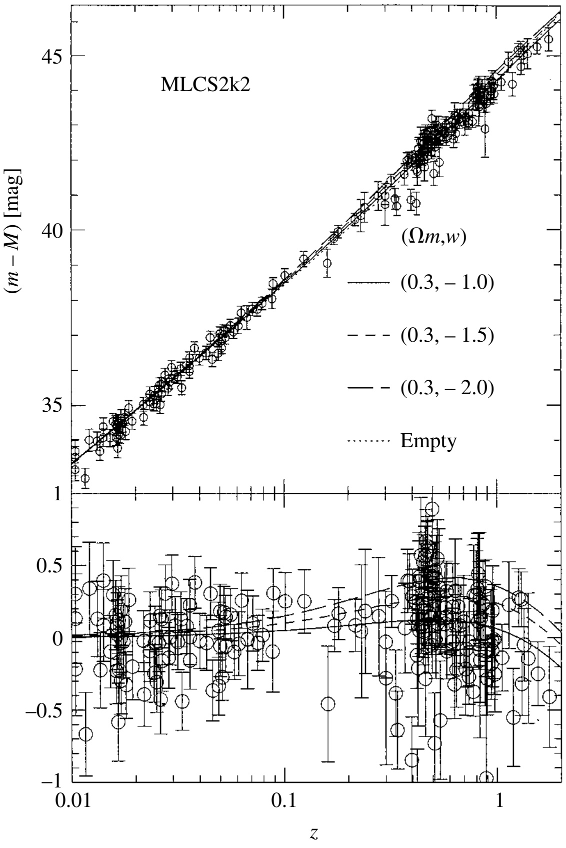


Fig. 7.13 Hubble plot from Type Ia supernovae at low and high redshifts, after Clocchiatti *et al.* (2006). The upper panel shows the measured values of the distance modulus (or logarithm of the luminosity distance) plotted against redshift. The lower panel shows the *difference* in magnitude as compared with the value expected for an empty universe. For averaged values, see Fig. 7.14.

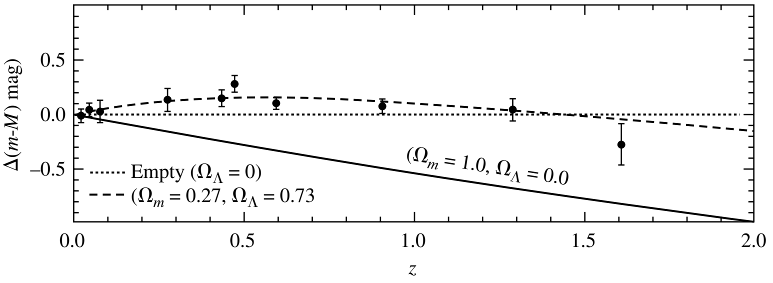


Fig. 7.14 Differential Hubble plot from Type Ia supernovae, after Riess *et al.* (2004). The experimental points represent averages over several supernovae. An empty universe ($\Omega_k = 1$, $\Omega_m = \Omega_\Lambda = 0$) is represented by the horizontal dotted line. A flat, matter-dominated (so-called Einstein–de Sitter) universe ($\Omega_m = 1$, $\Omega_\Lambda = \Omega_k = 0$) is shown by the solid curve; while the dashed curve represents their best fit, with $\Omega_m = 0.27$, $\Omega_\Lambda = 0.73$, $\Omega_k = 0$.

experimental results have been averaged, so that the trends are more easily visible. The dotted horizontal line again indicates an empty universe, the solid curve is for a matter-dominated universe, and the dashed curve corresponds in this case to a best fit with $\Omega_m = 0.27$, $\Omega_\Lambda = 0.73$. One again observes that the transition from an accelerating to a decelerating universe occurs at $z \sim 0.5$.

The results in Figs. 7.13 and 7.14 clearly exclude a flat, matter-dominated universe ($\Omega_m = 1$). The best fit to all available data today is for a flat universe with $\Omega_m = 0.24$, $\Omega_\Lambda = 0.76$, as given in (5.33) and (5.35). We repeat that these supernova results are in good agreement with the independent estimates from observations of large scale galaxy surveys, combined with analysis of anisotropies in the microwave background radiation described in Chapter 8, as well as with estimates from independent determinations of the age of the universe, for example, from globular clusters (see Example 5.3 and Section 10.3).

Example 7.3 *In a flat universe, with* $\Omega_m(0) = 0.24$ *and* $\Omega_\Lambda(0) = 0.76$*, at what value of z will the acceleration/deceleration be zero?*

From (5.46), $q(z) = \frac{1}{2}\Omega_m(0)(1+z)^3 - \Omega_\Lambda(0)$, which is zero when $(1+z) = [2\Omega_\Lambda(0)/\Omega_m(0)]^{1/3}$, or $z = 0.85$, somewhat larger than the z value of the flat maximum in the dotted curve in Fig. 7.14. A universe which is neither accelerating nor decelerating is often said to be 'coasting'.

We should note here that the analyses described above involve a comparison of high redshift with low redshift supernovae, so that the absolute luminosity scale, which is of course necessary in order to measure the Hubble parameter, is not required when comparing slopes. There are also potential complications in comparing the luminosities of supernovae at different redshifts, since they occur at different epochs and the metal content in early stars may be less than that in more recent ones, which have formed from the recycled debris of earlier stellar generations, and this could affect the opacity and hence the luminosity. This and other possible differential effects, such as dimming due to absorption or scattering by dust, have been analysed by the various research groups in exhaustive detail, and shown to have only minor effects on the results.

7.14.3 Interpretation of the supernovae Ia results

As indicated in Chapter 5, the acceleration of the universe in recent times ($z < 0.5$) has been interpreted in terms of a vacuum energy term, which is time and z-independent, and identical with Einstein's cosmological constant (see (5.23)). At present it is not at all clear that the dark energy is actually associated with such a vacuum state. For example, it could arise as the 'latent heat' as a result of some sort of phase transition. Or the dark energy could be the manifestation of some new type of evolving scalar field, in so-called 'quintessence' models. More radical suggestions—so far with absolutely no supporting evidence—are that the dark energy (or dark matter) terms have appeared artificially, because of subtle deviations from Newton's inverse square law of gravitational attraction at very large, cosmological distances, as discussed in Section 2.10. In fact there is compelling evidence for dark matter, shown above in Fig. 7.6, which is quite independent of assumptions about any possible deviations from the inverse square law.

Assuming that the dark energy is not an artefact, the equation of state describing it may of course be different from that for the vacuum in Table 5.2, and the ratio of pressure to density could be time-dependent. From the supernovae results, when combined with data from galaxy surveys and from the acoustic peak analysis of cosmic microwave background (CMB) fluctuations described in Chapter 8, it is possible to measure the quantity $w = P/\rho c^2$ occurring in the equation of state for the dark energy, but only by assuming that it is time-independent. Then one finds as in (5.36)

$$w_{\text{darkenergy}} = -0.97 \pm 0.08 \tag{7.24}$$

consistent with the value $w = -1$ for a simple vacuum/cosmological constant.

One of the major difficulties in identifying the dark energy with vacuum energy arises when one considers its time evolution. For example, the ratio of vacuum energy density to matter energy density today is $\rho_\Lambda/\rho_m \sim 3$, but while ρ_Λ is constant, $\rho_m \propto R^{-3} \propto (1+z)^3$. Thus $\rho_\Lambda/\rho_m \sim 3/(1+z)^3$, and at the time of decoupling of matter and radiation, when $(1+z) \sim 1100$, the ratio ρ_Λ/ρ_m would have been 10^{-9} only. Conversely, in the future the ratio will become very large, as the matter density falls off as $1/R^3$ and the scale parameter R will eventually increase exponentially with time.

Given this very large variation in the relative contribution of vacuum energy to the total closure parameter, a major puzzle is the fact that at the present epoch it just happens to be within a factor 3 of the matter energy density. To circumvent this problem, one can postulate that the dark energy is associated with some new type of scalar field—called *quintessence*—for which the equation of state is such that the ratio $w = P/\rho c^2 < -1/3$ so as to ensure an accelerated expansion (see (5.46)), and is time dependent and of magnitude to reproduce the value measured in (7.24) at modest z-values. The quintessence field can be arranged to have an energy density which at early times closely follows or *tracks* (but is less than) the density of radiation, and after the era of matter–radiation equality, tracks the matter density. Referring to Table 5.2, it is noted that both radiation and matter energy densities vary as $1/t^2$, so that if the quintessence field had this property, its energy density would be a constant fraction of the total energy density. Obviously small variations on the $1/t^2$ dependence are also possible, for not too complicated quintessence potentials.

Over the years, there have been many interesting suggestions regarding the origin of dark energy. Could it be somehow related to a breakdown of Newtonian gravity at very *short* distances, reminiscent of the 'curled up' extra dimensions in supergravity models? The magnitude of the dark energy density from Ω_Λ is

$$\varepsilon = 4.1\,\mathrm{GeV\,m^{-3}},$$

and if we express this in natural units ($\hbar = c = 1$), this density will correspond to a fundamental length given by

$$L^4 = \frac{\hbar c}{\varepsilon},$$

where $\hbar c = 0.197\,\mathrm{GeV\,fm}$ has dimensions of energy × length, and ε has dimensions of energy/(length)^3. One thus finds a value for L of 84 μm. Unfortunately for this proposal, the inverse square law has in fact been found to be valid using a very precise torsion balance, over the range from 9 mm down to 55 μm (Kapner *et al.* 2007).

7.15 Vacuum energy: the Casimir effect

In Chapter 5 we already mentioned that present observations, such as those described in the previous section, appear to require for their interpretation a major contribution from dark vacuum energy to the present energy density of the universe. The vacuum energy itself is postulated to arise through quantum fluctuations, that is, the spontaneous appearance and disappearance of virtual particle–antiparticle pairs and quanta, as required by the uncertainty principle. That this concept is not just a figment of the physicist's imagination was already demonstrated many years ago, when Casimir (1948) predicted that by modifying the boundary conditions on the vacuum state, the change in vacuum energy would lead to a measurable force, subsequently detected and measured by Spaarnay (1958) and more recently and comprehensively by Lamoreaux (1997) and Roy *et al.* (1999).

Essentially the Casimir effect in its original configuration arises when two perfectly conducting, flat parallel plates are placed close together with a very small separation a (see Fig. 7.15). The vacuum energy between the plates is different from that in the same volume with the plates absent, because the plates introduce boundary conditions on the fluctuating field. For example, if the virtual quanta are those of the electromagnetic field, there are boundary conditions on the associated electric and magnetic fields (the component of **E** parallel to the plates and of **B** normal to the plates must vanish at the surface, so that if the x-axis is normal to the plates, wavenumbers $k_x < \pi/a$ are forbidden). This difference in vacuum energy corresponds to a force between the plates which is actually attractive in this particular configuration (the sign of the force in general depends on the geometry).

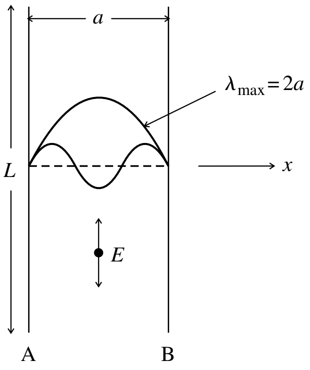


Fig. 7.15 End-on view of parallel plates A and B in an experiment to measure the Casimir effect, demonstrating the existence of vacuum energy density. The electric field E must vanish at the surface of perfectly conducting plates, so that $\lambda(\text{max}) = 2a$ or $k_x\,(\text{min}) = \pi/a$, and such boundary conditions change the value of the vacuum energy and give rise to an attractive force. Successful attempts to verify the effect have used the experimentally simpler configuration of a plate and a hemisphere rather than two plates.

Just on the basis of dimensional arguments, one can understand that the force per unit area in Fig. 7.15 must be of order $\hbar c/a^4$. Planck's constant times the velocity of light must enter, as it does in all uncertainty relation problems, and gives dimensions of energy times length, which has to be divided by the fourth power of a length in order to get a force per unit area. If the plates are of side $L \gg a$, the only length of relevance is the separation a. We simply quote here the result of a full calculation (see, for example, Itzykson and Zuber 1985).

$$F = -\frac{\left(\pi^2/240\right)\hbar c}{a^4} \sim \frac{13}{a^4}\,\mu\text{g wt cm}^{-2} \qquad (7.25)$$

where the plate separation a is in microns (μm). This tiny force, of order micrograms weight per square centimetre, and its dependence on plate separation, has been measured and the above formula verified to within 1% accuracy. Of course, the effect does not measure the absolute value of the vacuum energy density, but only the change when the topology is altered. On the other hand, the gravitational field couples to the *absolute* values of energy and momentum, and the total vacuum energy can only be measured *via* its gravitational effect.

The Casimir effect has implications outside quantum field theory and cosmology, for example, in electromechanical systems on submicron scales, where it could lead to malfunctions of the system. There are also classical macroscopic analogues of the Casimir effect. The most famous is known to all mariners. Under certain wave conditions, two ships sailing close together beam-to-beam experience a force of attraction, due to the fact that the wave pattern between the ships is affected by the presence of the hulls and certain wavelengths are again suppressed (see Buks and Roukes (2002) for reference).

7.16 Problems with the cosmological constant and dark energy

The cosmological constant $\Lambda = 8\pi\mathrm{G}\ \rho_{\mathrm{vac}}$ presents one of the major—if not *the* major—conceptual problems in cosmology, and has done so ever since Einstein introduced it. It has long been argued that the dark energy density associated with the cosmological constant ought to have a 'natural' value determined by the scale of gravity. This natural unit is then the Planck mass energy $M_{\mathrm{PL}}c^2 = \left(\hbar c^5/\mathrm{G}\right)^{1/2} = 1.2 \times 10^{19}$ GeV, in a cube of side equal to the Planck length $\hbar/M_{\mathrm{PL}}c$ (see (1.12)), that is, an energy density

$$\frac{\left(M_{\mathrm{PL}}c^2\right)^4}{(\hbar c)^3} \sim 10^{123}\,\mathrm{GeV\,m^{-3}} \tag{7.26}$$

a truly gigantic number, which of course is nonsensical since it would imply that the universe could only be a few seconds old at most. So, it is perhaps instructive to see in more detail how this number is arrived at.

In quantum field theory one can describe the vacuum fluctuations of the boson fields as due to an ensemble of simple harmonic oscillators of different frequencies. The energy of one such (bosonic) oscillator is $(n + 1/2)\,\hbar\omega$ where ω is an angular frequency and $n = 0, 1, 2, \ldots$. The vacuum or ground state has the so-called 'zero-point energy' $E = \frac{1}{2}\hbar\omega$. In a sense, it is a matter of choice whether one takes this zero-point energy seriously or simply ignores it, since measurements are usually about energy differences, and only when we come to gravity do we have to worry about the absolute energy value. However, if we try to identify it with the mysterious, dark vacuum energy, then we have to sum over all oscillators in the volume. From (1.16), the number of possible quantum states in a spatial volume V, with wave numbers $k = p/\hbar$ lying in the element $k \to k + \mathrm{d}k$ and integrated over all directions, is $4\pi V\,k^2\,\mathrm{d}k/(2\pi)^3$. So the total energy per unit volume of all the oscillators will be

$$\varepsilon = \frac{E}{V} = \left[\frac{\hbar}{\left(4\pi^2\right)}\right] \times \int k^2 \mathrm{d}k\ \omega_k \tag{7.27}$$

The angular frequency is related to the wavenumber by $\omega_k^2 = k^2c^2 + m^2c^4/\hbar^2$ where m is the oscillator mass. Obviously this integral is divergent, but let us cut it off at some value k_m or $E_m \gg mc^2$. Then, with $\omega_k \approx kc$ in the relativistic approximation

$$\varepsilon = \left(\frac{\hbar c}{16\pi^2}\right) k_m^4 = \frac{E_m^4}{\left[16\pi^2\,(\hbar c)^3\right]} \tag{7.28}$$

Here, the cut-off is arbitrary. For example, we can place it at an energy scale where we expect quantum field theory to start to fail, and that is the Planck scale $E_m = M_{PL}$ of quantum gravity. Another excuse for choosing this scale is that it is the 'natural' scale of energy from combining the fundamental constants G, $\hbar$, and c, that is, $(\hbar c/\mathrm{G})^{1/2}$. Including the numerical constants left out in (7.26), this gives $\varepsilon \sim 10^{121}$ GeV m^{-3} as before, to be compared with the critical energy density in (5.26) of $\rho_c \sim 5$ GeV m^{-3}, of which only a part can be assigned as vacuum energy. So why is the observed vacuum energy/cosmological constant only about 10^{-121} of the naïve expectation?

Of course, one might vary the value of E_m so that the vacuum energy density is of the same order as the critical density, that is, $\varepsilon = \rho_c = 5$ GeV m^{-3}. Then substituting in (7.28) would give $E_m \sim 0.01$ eV only. This is ridiculously small in comparison with the masses of practically all known elementary particles, or even of atomic energy levels. It has been remarked, however, that this energy is indeed comparable with mass differences of light neutrinos (4.12). It would indeed be extraordinary if the acceleration of the universe were somehow tied up with neutrino masses, but the mystery of dark energy is such that even the most outlandish ideas cannot be ignored.

Twenty-five years ago, before the importance of dark energy was fully apparent, it was believed that the matter density was such that $\Omega_m \sim \Omega_{tot} \sim 1$ and that the cosmological constant might even be identically zero, and dark matter made up the vast bulk of the energy density. In view of the above argument, the difficulty then was to understand why the cosmological constant was so incredibly small, or even zero. Here, at least one can say that zero is a natural number, for which a reason might be found. For example, the masslessness of the photon is associated with a symmetry principle, namely the local gauge invariance of the electromagnetic interaction as described in Section 3.7. However, no symmetry principle is known that could set $\Lambda = 0$. Indeed, the finiteness of the vacuum energy/cosmological constant seems to follow inexorably from quantum mechanics, for the very simple reason that the virtual states of the real particles which contribute to Ω_m *must* contribute to Ω_Λ.

However, the above integral will include the summed effects of all types of elementary field, with somewhat different amplitudes and phases, and there can be cancellations. For example, the energy of a fermion oscillator, analogous to the above expression for a boson oscillator, is $(n - 1/2)\,\hbar\omega$, so that the zero-point energy comes in with the opposite sign. (This comes about because the wave functions describing creation and annihilation of bosons and fermions obey commutation and anti-commutation relations respectively.) So in a theory of *exact* supersymmetry, where every boson is matched by a fermion of the same mass and vice-versa, there would indeed be complete cancellations, with a vacuum energy of zero. However, we know that in the real world, even if supersymmetry turns out to be valid, it must be a badly broken symmetry. While at very large values of k in (7.27), well above the supersymmetric scale, there could be exact cancellations, this would not be the case at lower k values.

The actual situation is of course somewhat worse than this, since the supernovae results present us with a finite, non-zero number for the dark energy density, inconceivably small in comparison with what might be expected, but one for which the relative contribution to the overall energy density apparently changes with time. Attempts to model this behaviour have been described in the previous section.

Finally, a different approach to the problem has been to appeal to the anthropic principle, namely that life exists only when the laws of physics allow it. In this case, it is the value for Λ at the present epoch. Had it differed by just an order of magnitude or so, there would have been no human race to ponder on the problem. As the saying goes, we live in the best of all possible worlds. This argument perhaps becomes more plausible in the context of inflationary models of the early universe, described in Chapter 8. These suggest that our particular universe is just one of an enormous number of parallel universes, so that the human race could have evolved in the one where conditions happened to suit it.

In summary, the phenomenon of the cosmological constant or dark energy, accounting at the present time we believe for the bulk of the energy in the universe, is simply not understood, and this, like our incomprehension of the matter–antimatter asymmetry of the universe, could be ranked as a major failure in the subjects of cosmology and particle physics. These failures have not grown up overnight. The problems of dark matter and of the vacuum energy/cosmological constant have been lurking for at least 70 years, but they have become more acute in the last two decades because of the vastly improved quality and quantity of the experimental data, and the remarkable discovery of an accelerating universe. However, it cannot be too strongly emphasized that it is *precisely* such problems which keep the subject of particle astrophysics alive and stimulating and full of great challenges for the future.

7.17 Summary

- The rotation curves of stars in spiral galaxies imply that the bulk of the matter (80–90%) is non-luminous, and located in a galactic halo.
- Studies of X-rays from galactic clusters indicates velocities of the gas particles emitting the X-rays which are far in excess of escape velocities based on the visible mass.
- Dark matter is also required in cosmological models of the early universe, if the structure of galaxies and galaxy clusters is to evolve from the very small primordial density fluctuations deduced from inhomogeneities in the microwave radiation (discussed in Chapter 5).
- Independent evidence for dark matter is found from the gravitational lensing of distant galaxies and clusters by foreground galaxies. Gravitational microlensing of individual stars, appearing as a temporary, achromatic enhancement in luminosity, shows that some of the galactic dark matter is baryonic, this matter appearing in the form of so-called MACHOs, which are dark star-like objects with masses of $0.001 - 0.1$ solar masses.
- Baryonic dark matter makes less than 25% contribution to the total dark matter density, and the bulk of dark matter is non-baryonic.
- The most likely candidates for dark matter are WIMPs, that is, very massive, weakly interacting particles, constituting ‘cold’ dark matter. Until the nature of such particles is established, the most common suggestion is that they are supersymmetric particles such as neutralinos.
- Several experiments are under way to detect WIMPs directly by observing nuclear recoils from elastic scattering of WIMPs, so far with no success.
- Observations of Type Ia supernovae at high redshifts ($z \sim 1$) suggest that in earlier times the expansion rate (i.e. the Hubble parameter H) was less than it is today; or that, relative to earlier times, the universal expansion is now accelerating.

- The present acceleration is interpreted in terms of a finite value for the cosmological constant, or for the existence of dark (vacuum) energy. This dark energy seems to account for some 2/3 of the total energy density of the universe today.
- The reality of vacuum energy is evidenced by the laboratory observation of the Casimir effect, which is a manifestation of a change in the vacuum energy when boundary conditions are imposed on it.
- There is no satisfactory explanation for the observed magnitude of the dark energy, which at the present time is comparable in magnitude with the matter energy density. If the dark energy is identified with vacuum energy, then in times past it would have been negligible while in the future it will become dominant. Other possible sources of dark energy have been proposed, involving completely new types of interaction. However, the puzzle of the present magnitude of the dark energy remains.

Problems

More challenging or longer questions are denoted by an asterisk.

(7.1) Estimate the angular deflection of a photon by a point mass M, according to Newtonian mechanics. Express the result in terms of b, the closest distance of approach.

(7.2) Calculate an expression for the tangential velocity v of a star near the edge of the disc of a spiral galaxy of radius R and mass M, and thus find an expression for the optical depth τ for microlensing in terms of v. Give numbers for the Milky Way, with a mass of 1.5×10^{11} solar masses and a disc radius of 15 kpc.

(7.3) Obtain an expression for the kinetic energy E_{R} of a nucleus of mass M_{R} recoiling in an elastic collision with a dark matter particle of mass M_{D} and incident kinetic energy E_{D}, in terms of the angle of emission relative to the incident direction. Find the limiting values of recoil energy in terms of M_{D} and M_{R}. Calculate the maximum recoil energy of a nucleus of 80 proton masses, in collision with a dark matter particle of mass 1000 times the proton mass, travelling with a typical galactic velocity of 200 km s^{-1}.

(7.4) Assume that the universe is flat, with $\Omega_m(0) = 0.24$, $\Omega_\Lambda(0) = 0.76$. What is the numericalvalue of the acceleration or deceleration with respect to the Earth, of a galaxy at redshift $z = 0.03$? Compare this with the local acceleration (g) due to the Earth's gravity. Neglect the 'peculiar velocity' of the Earth with respect to the Hubble flow and assume $H_0 = 70$ km s^{-1} Mpc^{-1}.

(7.5) Show that, if the vacuum energy density and matter energy density today are comparable in magnitude, then when the universe was a fraction f of its present age, the relative contribution of the vacuum energy would have been f^2.

(7.6) Verify the results given in Table 7.1 for luminosity distances as a function of redshift.

8 Development of structure in the early universe

8.1 Preamble

The Big Bang model described in Chapter 5 seems to give a rather convincing description of the development of the early universe. It is underpinned by three striking phenomena:

- The observation of the redshift of distant galaxies.
- The correct prediction of the abundances of the light elements via primordial nucleosynthesis.
- The existence of the all-pervading cosmic microwave background (CMB) radiation.

This success is all the more remarkable since the principal tenets of the model—isotropy and homogeneity of the 'cosmic fluid'—are to be contrasted with the universe today, characterized by a decidedly non-isotropic, non-homogeneous nature—galaxies, galactic clusters, voids, and so forth. The question arises: how did we get from the uniformity of the Big Bang model to the present universe with its lumpy structure?

As described in the following pages, it can plausibly be argued that this structure had its origins in quantum fluctuations in energy density which occurred in the very early universe and were then 'frozen-out' when the universe underwent an exponential and superluminal expansion stage called *inflation*. These tiny fluctuations in density and temperature—typically at the 10^{-5} level—then acted as seeds for the development of much greater fluctuations in density via the subsequent process of gravitational collapse during the epoch of matter domination.

In Section 8.3 we outline the inflation scenario, which was postulated over two decades ago in response to some difficulties with the Big Bang model, mainly with respect to the initial conditions which are apparently required. Later in the chapter we touch on the subject of galaxy formation. One of the most remarkable features of the universe is that stars are always clumped into galaxies each containing of order 10^{11} stars. The galaxies are separated by distances some two orders of magnitude larger than their diameters ($\sim$ Mpc as compared with $\sim$ 10 kpc), and one might ask the question; why has matter become distributed in this particular fashion—rather than, for example, in one giant galaxy? To anticipate our discussion, the present answer appears to be that in the early universe primordial density fluctuations could only start to grow, provided that they were spread out over distances which now correspond to the dimensions of galaxy clusters, and that these dimensions were in turn determined by the properties and interactions of the primordial photons and neutrinos which dominated the radiation era.

In this text, we do not enter into the very complex cosmology of the distribution and formation of galaxies, but only discuss in general terms how the universe could have progressed from the uniformity of the Big Bang model to the very lumpy structure of the universe today.

8.2 Galactic and intergalactic magnetic fields

In this section we discuss very briefly the nature and magnitude of the intergalactic magnetic fields, to try to assess whether they were important in the early development of the universe. Inside our own galaxy, the interstellar magnetic fields are very significant and follow the spiral arms. The average value of this galactic field B $\sim 3\mu$G (0.3 nT) and its energy density is therefore $B^2/8\pi \sim 0.2$ eV cm^{-3}, so is comparable with the energy density of cosmic microwave radiation (0.26 eV cm^{-3}) and that of the cosmic rays in deep space ($\sim$1 eV cm^{-3}). This galactic field has been measured, for example, by observing the Faraday rotation of polarized light from pulsars (Han *et al.* 2006). The rotation is proportional to the line integral of the B-field and the square of the wavelength, so the mean field can be found from the wavelength dependence of the rotation, and varies from about 2×10^{-6} G in outer parts of the spiral arms to 4×10^{-6} G towards the central hub.

The intergalactic field is known only within limits and is certainly very much less than that inside the galaxy. Magnetic fields associated with galaxy and cluster filaments, that is, on the scale of 1 Mpc or less, have been estimated from the soft synchrotron radiation emitted by electrons traversing the field, and are of order 10^{-7} G or less (Kronberg 2004). For very deep space (i.e. distances of order 10–100 Mpc), a value of 10^{-11} G is indicated from the observation in the AUGER experiment (see Section 9.13) that extensive air showers produced by protons of energy above 6×10^{19}eV are correlated (within about 3^0) with known point sources (active galactic nuclei, AGNs) up to 75 Mpc distant (Dermer 2007).

Some models postulate that very weak and diffuse intergalactic fields are produced by the passage of high-energy charged particles—cosmic rays—produced as ejecta and jets from supernova explosions, that is, following the era of star formation (roughly, for redshifts $z < 12$). Another view is that the fields could have been generated by dynamic amplification of weaker ‘seed’ fields generated much earlier, in the primordial proton–electron plasma before the era of recombination (i.e. for $z > 1100$). The density fluctuations in the plasma would produce ‘winds’ of photons streaming from high to low density regions. The idea is that such photons would interact to separate the lighter electrons from the heavier protons, giving rise to charge separation and rotating electric currents seeding primordial magnetic fields.

However, since the present intergalactic field appears to be less than 10^{-5} of the galactic field, it seems unlikely that magnetic fields could have played a very significant role in the development of structure on very large (cluster and supercluster) scales. We shall follow the conventional wisdom, which treats gravity as the main player in these developments, with electromagnetic interactions in a subsidiary role.

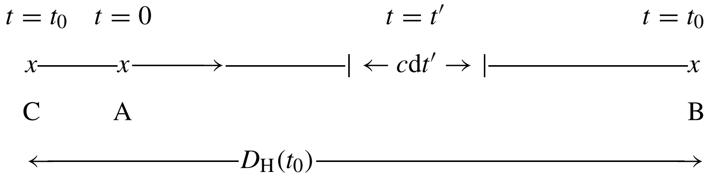
Of course it should be emphasized here that on smaller, stellar scales, magnetic fields are of extreme importance. For example, as discussed in Chapters 9 and 10, the very intense fields associated with supernova explosions are considered to be the main accelerators of the cosmic rays, in which particle energies extend to 10^{15} eV and beyond. The thousands of pulsars which have been observed are constant reminders of the enormous magnetic fields which are generated in the later stages of massive stars.

8.3 Horizon and flatness problems

We first discuss two of the principal problems of the Big Bang model, with regard to the initial conditions required. These are known as the horizon and the flatness problems.

The *particle horizon* is defined as the distance out to which one can observe a particle, by exchange of a light signal. In other words, the horizon and the observer are causally connected. More distant particles are not observed, they are beyond the horizon. The horizon is finite because of the finiteness of the velocity of light and the finite age of the universe. In a static universe of age t, we expect to be able to observe particles out to a horizon distance $D_H = ct$. As time passes, D_H will increase and more particles will move inside the horizon. At the present time, the universe has age $t_0 \sim 1/H_0$—see Section 5.6. The quantity $ct_0 \sim c/H_0$ is usually referred to as the *Hubble radius*, that is, the product of the Hubble time and the velocity of light.

In an expanding universe, it is apparent that the horizon distance will be somewhat greater than ct. Let us assume, as appears to be the actual case, that we are dealing with a flat universe with zero curvature ($k = 0$), as indicated by the measurements described later in this chapter, and hence that on extremely large scales, light travels in straight lines Suppose that a light signal leaves a point A at $t = 0$ (see figure below) and arrives at the point B at $t = t_0$. By the time $t = t_0$, A will have moved, relative to B, to the point C.



Consider the time interval $\mathrm{d}t'$ at time t', where $0 < t' < t_0$. The light signal will cover a distance $c\mathrm{d}t'$, but because of the Hubble expansion, by the time $t = t_0$, this will have expanded to $c\mathrm{d}t'R(0)/R(t')$, where $R(t)$ is the scale parameter in (5.8) and $R(0)$ is its present value. Hence the horizon distance will be

$$D_H(t_0) = R(0) \int \frac{c\mathrm{d}t}{R(t)} \tag{8.1}$$

integrated from $t = 0$ to $t = t_0$. Since from (5.9)

$$R(t) = \frac{R(0)}{(1+z)},$$

then

$$\mathrm{d}t = \frac{\mathrm{d}z}{[(1\text{+z})\,H]},$$

so that

$$\frac{R(0)c\mathrm{d}t}{R(t)} = \frac{c\mathrm{d}z}{H} \tag{8.2}$$

Therefore the horizon distance today, expressed in terms of the redshift z is

$$D_{\mathrm{H}}(t_0) = c\int \frac{\mathrm{d}z}{H} \tag{8.3}$$

where from (5.37)

$$H = H_0\left[\Omega_m(0)(1+z)^3 + \Omega_r(0)(1+z)^4 + \Omega_\Lambda(0) + \Omega_k(0)(1+z)^2\right]^{1/2}$$

The fraction of the universe, of current dimension $R(0)$, which is inside the optical horizon today is therefore proportional to

$$F = \frac{D_{\mathrm{H}}(t_0)}{R(0)} = \left[\frac{c}{H_0R(0)}\right] \times \int \frac{\mathrm{d}z}{\left[\Omega_m(0)\,(1+z)^3 + \Omega_r(0)\,(1+z)^4 + \Omega_\Lambda(0) + \Omega_k(0)(1+z)^2\right]^{1/2}} \tag{8.4}$$

integrated from $z = 0$ to $z = \infty$. We see from the integrand that if the lower limit of the integral is taken as z^*, then for a matter-dominated universe, F decreases as $1/\sqrt{(1+z^*)}$ or $t^{1/3}$, while it falls off as $1/(1+z^*)$ or $t^{1/2}$ in the case of radiation dominance (see Table 5.2). Thus at early times or large z values, the fraction of the universe inside the horizon was much smaller than it is now.

This result also follows more briefly by observing that for most cosmological models, $R(t) \propto t^n$ where $n < 1$, so that integrating from $t = 0$ to $t = t_0$ the above formula gives

$$D_{\mathrm{H}}(t_0) = \frac{ct_0}{(1-n)} \tag{8.5}$$

One observes that the ratio

$$\frac{D_{\mathrm{H}}(t)}{R(t)} \propto t^{(1-n)} \tag{8.6}$$

so that since $n < 1$, again the fraction of the universe which is causally connected was once much smaller than it is now.

Example 8.1 *Calculate the particle horizon distance for a flat universe dominated by (a) matter and (b) radiation.*

Refer to Table 5.2. For a matter-dominated universe $n = 2/3$ and hence $D_{\mathrm{H}} = 3ct_0 = 2c/H_0$; while for the case of radiation domination, $n = 1/2$ and $D_{\mathrm{H}} = 2ct_0 = c/H_0$, where we have used the fact that the ages for the matter- and radiation-dominated cases are $t_0 = 2/(3H_0)$ and $1/(2H_0)$ respectively. With $1/H_0 = 14$ Gyr, the corresponding horizon distances are 2.5×10^{26} and 1.25×10^{26} m respectively.

In particular, the time of decoupling of matter and radiation was $t_{\mathrm{dec}} = 4 \times 10^5$ years (see Section (5.12)), and the horizon size then would have been approximately ct_{dec}. By now, this would have expanded to $ct_{\mathrm{dec}}(1+z_{\mathrm{dec}})$ where from (5.75) $z_{\mathrm{dec}} = 1100$. Hence, the angle subtended by that horizon distance at the Earth today for the case of a flat, matter-dominated universe would be, with $t_0 = 1.4 \times 10^{10}$ years

$$\theta_{\mathrm{dec}} \sim \frac{ct_{\mathrm{dec}}\,(1+z_{\mathrm{dec}})}{2c\,(t_0 - t_{\mathrm{dec}})} \sim 1^\circ \tag{8.7}$$

This formula shows that only the microwave radiation observed over small angular scales, of order one degree or so, corresponding to the time of the last interaction of these photons, could ever have been causally connected and in thermal equilibrium with other matter. On the contrary, after allowing for a dipole anisotropy associated with the 'peculiar velocity' of the Earth with respect to the microwave radiation, the temperature of the radiation is found to be uniform to within one part in 10^5, out to the very largest angular scales. This is the horizon problem.

The *flatness problem* arises as follows. From (5.26) and (5.27) the fractional difference between the actual density and the critical density is

$$\frac{\Delta\rho}{\rho} = \frac{(\rho - \rho_c)}{\rho} = \frac{3kc^2}{8\pi G R^2\,\rho} \tag{8.8}$$

During the radiation-dominated era, $\rho \propto R^{-4}$. From (8.8) it follows that $\Delta\rho/\rho \propto R^2 \propto t$. So at early times, $\Delta\rho/\rho$ must have been much smaller than it is today, when $t \sim 4 \times 10^{17}$ s and it is of order unity. For example, for $kT \sim 10^{14}$ GeV, a typical energy scale of grand unification, $t \sim 10^{-34}$ s, and at that time $\Delta\rho/\rho$ would have been $\sim 10^{-34}/10^{18} \sim 10^{-52}$ (and even smaller than this if we include the period of matter dominance). How then could $\Omega = \rho/\rho_c$ have been so closely tuned as to give of the order of unity today?

In short, these two problems require a mechanism which allows thermal equilibrium outside conventional particle horizons, and can reduce the curvature k/R^2 in (8.8) by a huge factor. A possible answer—indeed, the only one we have—was supplied by Guth in 1981 (who was actually concerned to reduce the possible monopole flux—see below). He postulated an extremely rapid exponential expansion by a huge factor as a preliminary stage of the Big Bang, a phenomenon known as *inflation.* Since that time, there have been a number of inflationary models—old inflation, new inflation, chaotic inflation, eternal inflation, and so on—none of them yet fully capable of a completely satisfactory description. However, there seems to be little doubt that some sort of inflationary scenario is an obligatory first stage of the birth pangs of the universe.

8.4 Inflation

In this section we give a brief and qualitative description of the inflation scenario. First we recall the Friedmann equation (5.11)

$$\left(\frac{\dot{R}}{R}\right)^2 = \frac{8\pi G\,(\rho + \rho_\Lambda)}{3} - \frac{kc^2}{R^2}$$

where $\rho = \rho_m + \rho_r$ is the energy density of matter and radiation, and ρ_Λ is the vacuum energy density, which as explained in Chapter 5 is a space- and time-independent quantity. Suppose a situation arises in which ρ_Λ dominates the other terms on the right-hand side of the equation. Then the fractional expansion rate becomes constant and one obtains *exponential growth* over some time interval between t_1 when inflation commences and t_2 when it terminates:

$$\left(\frac{\dot{R}}{R}\right)^2 = \frac{8\pi G\rho_\Lambda}{3}$$

and

$$R_2 = R_1 \exp\left[H(t_2 - t_1)\right] \tag{8.9}$$

where

$$H = \left(\frac{8\pi G\rho_\Lambda}{3}\right)^{1/2} = H_0\left(\frac{\rho_\Lambda}{\rho_c}\right)^{1/2}$$

Also, since RT is constant, the temperature will fall exponentially during the inflation era, that is, the energy per particle is red-shifted away by the expansion:

$$T_2 = T_1 \exp\left[-H\,(t_2 - t_1)\right] \tag{8.10}$$

As stated above, the horizon distance at the likely timescale ($t = 10^{-34}$ s) of the GUT (grand unified theory) era, for example, was $ct \sim 10^{-26}$ m. If we take as the notional size of the present day universe, when $t_0 \sim 4 \times 10^{17}$ s, the value $ct_0 \sim 10^{26}$ m, the radius at $t = 10^{-34}$ s would have been $\left(10^{-34}/4 \times 10^{17}\right)^{1/2} \times 10^{26} \sim 1$ m, that is, enormously larger than the horizon distance at that time. However, in the inflationary scenario, the physical size of the universe before inflation is postulated to be smaller than the horizon distance, so that there was time to achieve thermal equilibrium by causal interactions, which can take place over time intervals entirely dictated by the speed of light. During the inflationary period this tiny region has to expand and encompass the 1 m size of the universe which commences the conventional Big Bang 'slow' expansion, with $R \propto t^{1/2}$. This evolution therefore requires that

$$\exp H\,(t_2 - t_1) > 10^{26}$$

or

$$H\,(t_2 - t_1) > 60 \tag{8.11}$$

If this condition can be achieved, the horizon problem disappears, since even the most distant parts of the universe would once have been in close thermal contact, and it was only the superluminal expansion of space, far above the speed of light, which necessarily left them disconnected. The flatness problem is also taken care of, since the curvature term in (5.11) is reduced by a factor

$$\left(\frac{R_2}{R_1}\right)^2 = \exp\left[2H\,(t_2 - t_1)\right] \sim 10^{52}$$

so that if $\Omega(t_1)$ is only of the *order* of unity at the beginning of inflation, at the end of inflation it will be incredibly close to unity:

$$\Omega(t_2) = 1 \pm 10^{-52} \tag{8.12}$$

and on large enough, supergalactic scales the universe should be equally flat and uniform at the present day. An analogy can be made with the inflation of a rubber balloon: as it inflates, the curvature of the surface decreases and in the limit a portion of the surface appears quite flat.

There is one other problem solved by inflation, indeed it was the original motivation for Guth's model. *Magnetic monopoles* were suggested by Dirac in 1932, and are definitely predicted to exist in grand unified theories (where quantization of electric charge and therefore of magnetic charge appears naturally). The monopole masses would be of the order of the GUT mass scale and they should then have been created in the early universe with a number density comparable with photons. They would have survived as stable particles and their energy density would have dominated the universe. Searches for magnetic monopoles have met with no success and the observed upper limit on monopole density is many orders of magnitude below the above expectation. Provided however that monopoles, because of their large masses, can only be created at very high temperatures, *before* the inflationary process commences, the monopole problem is also solved, since the monopole number density will fall by an exponential factor through inflation and typically there would be only one monopole left in our entire universe. After inflation, the temperature is assumed to be too low to lead to monopole creation, but would of course be sufficient to create all the elementary particles we are familiar with in accelerator experiments.

What is the physical mechanism underlying inflation? That is unknown. It is simply postulated that it is due to some form of scalar field, called the *inflaton* field. In its original form, the inflation mechanism was likened to the Higgs mechanism of the self-interactions of a scalar field and spontaneous symmetry breaking in the very successful theory of the electroweak interaction described in Chapter 3, only at a much higher-energy scale, for example, the GUT scale, of perhaps 10^{14}–10^{16} GeV.

Suppose that one were to start off with an intensely hot microscopic universe near the Planck temperature $kT \sim 10^{19}$ GeV, expanding and cooling as in (5.49), and that the initial evolution suddenly became dominated at $t = t_1$ by such an 'inflaton' field ϕ, consisting of scalar particles of mass m. For $kT \gg mc^2$, the field is assumed to be in the ground state with a vacuum expectation value $\langle\phi\rangle = 0$ as in Fig. 8.1. This state is referred to as the 'false vacuum' state. At temperatures below a critical value $kT_c \sim mc^2$ however, through a process of spontaneous symmetry breaking, the vacuum expectation value of the field can become different from zero, with $\langle\phi\rangle = \phi_{\min}$ and a lower potential energy. The system will therefore try to make the transition from the metastable state of the 'false' vacuum to the 'true' vacuum. The inflationary phase occurs while the system is in the false vacuum state, during the period $t_1 \to t_2$ when the energy density is

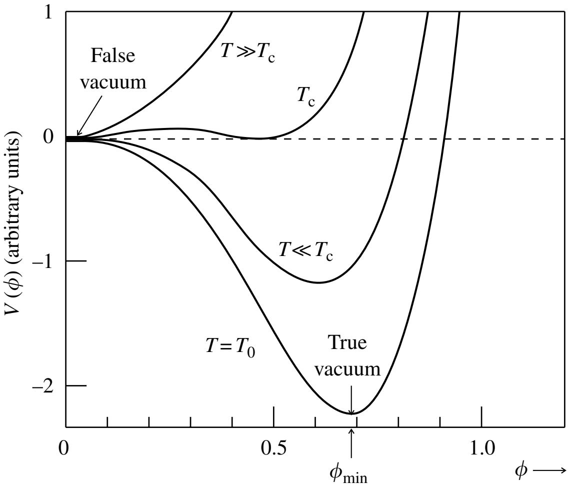


Fig. 8.1 Potential $V(\phi)$ of the 'inflaton' field plotted against the field vacuum expectation value $\langle\phi\rangle$ at different temperatures, in the early model of inflation. The critical temperature is denoted T_c. For temperatures slightly less than this value, a transition can be made from the 'false' to the 'true' vacuum *via* quantum-mechanical tunnelling. Inflation takes place while the system is in the 'false' vacuum state, and ends when it reaches the 'true' vacuum.

approximately constant. The inflationary expansion is of course driven by the vacuum energy.

It will be recalled that in Chapters 5 and 7 we saw that a large fraction (75%) of the energy density in the universe today is in the form of dark energy, perhaps as vacuum energy which is independent of the temperature (see Table 5.2). In the distant past this would have been a vanishingly small fraction, since the energy densities of radiation and of non-relativistic matter vary as T^4 and as T^3 respectively. In the context of inflation discussed here, we are calling upon a *second* and quite separate source of vacuum energy, existing only on an enormously high-energy scale and disappearing as soon as the inflationary stage is completed.

The inflationary phase is terminated in this model when the transition to the true vacuum occurs on account of quantum mechanical tunnelling through the potential barrier between the false and the true vacuum. 'Bubbles' of the true vacuum then develop, and these are supposed to merge into each other and stop the inflation. The energy density $\rho \sim (mc^2)^4$ which is liberated as the inflation ends and the system enters the true vacuum state, is the 'latent heat' which reheats the supercooled inflationary universe, so that it reverts to the conventional Big Bang model with 'slow' expansion and cooling. This reheating is analogous to the heat liberated when supercooled water suddenly undergoes a first-order phase transition to form ice, the supercooled water being the analogue of the false vacuum and the ice that of the true vacuum. The variations of R and T with time in this model are sketched in Fig. 8.2.

We already noted at the beginning of Chapter 5 that the gravitational potential of the universe today is almost equal to its mass energy, so the total energy is near zero. It is important to emphasize that, in the inflationary scenario, the universe starts out essentially from nothing, with zero total energy, as in equation (5.13) for a flat universe with $k = 0$. As the inflation proceeds, more and more positive energy appears in the rapidly expanding region occupied by the scalar field ϕ: eventually, after the transition to the true vacuum, the 'reheating' phase will lead to the creation of the enormous numbers of particles ($\sim 10^{88}$!) which eventually form the material universe. As this is happening, more and more

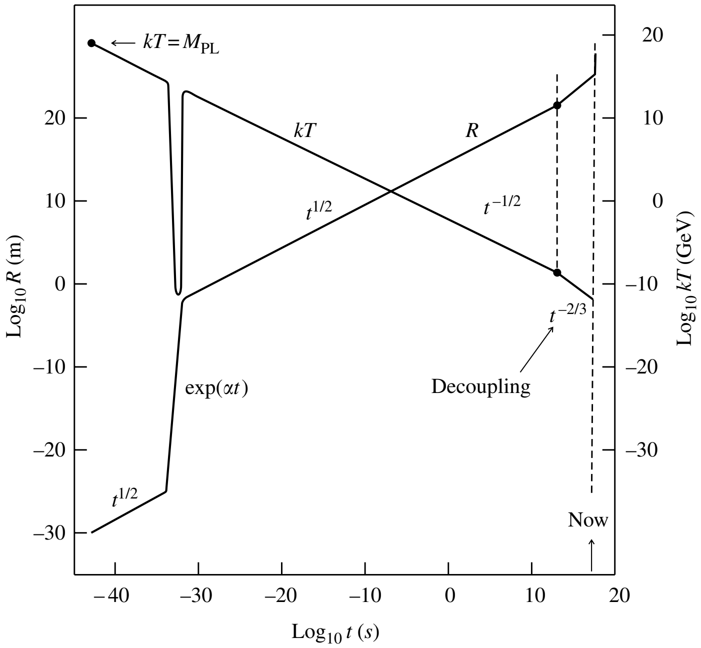


Fig. 8.2 Sketch of the variations of R and T with time in the inflationary scenario.

negative energy appears in the form of the gravitational potential energy of the expanding region. The total energy remains at a small and possibly zero value, with $k \approx 0$. The enormous energy associated with the expansion and particle creation is then simply provided by the gravitational potential energy of the expanding material. It is a bit like cycling down a hill, starting from rest at the top. The large kinetic energy acquired on reaching the bottom is exactly offset by the loss of potential energy due to the change in height.

The early Guth model of inflation sketched above suffered because it did not seem possible to obtain the necessary inflationary growth as well as to terminate the inflation efficiently so as to end up with a reasonably homogeneous universe. Wherever the transition between ‘false’ and ‘true’ vacuum takes place *via* quantum-mechanical tunnelling, ‘bubbles’ of true vacuum form and inflation ends. These bubbles will then grow slowly *via* causal processes, whereas outside them, exponential inflation continues, and one ends up with a very lumpy situation.

8.5 Chaotic inflation

The above problems are avoided in the *chaotic inflation* model, due originally to Linde (1982, 1984) and also Albrecht and Steinhardt (1982). In the Guth model described above, it is tacitly assumed that the universe commenced inflation when the inflaton field was exactly at the false vacuum minimum ($\phi = 0$). Linde pointed out that, because of quantum fluctuations at or near the Planck time, this was improbable, and the start value could be random. The basic idea is that, due to such fluctuations, conditions in different parts of the space–time domain vary in an unpredictable fashion, so that some regions attain the condition of inflation before others, and each such ‘bubble’ or ‘patch’ becomes a universe

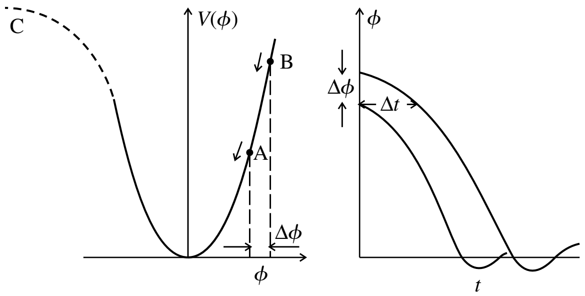


Fig. 8.3 A typical potential curve (on left) for the chaotic inflationary scenario, in this case for the quadratic potential (8.18). At large V, the curve must flatten out, as shown by the dashed line C, implied by the 'slow roll' approximation. No phase transition is involved in this case. Quantum fluctuations in ϕ mean that different points in the universe, say at A and B, begin and end inflation at different times (see right-hand diagram), separated by an interval $\Delta t = \Delta\phi/\dot{\phi}$ as in (8.22).

on its own. The inflaton potential is assumed to be a smooth function as in Fig. 8.3 (in this case the quadratic function in (8.18)). No phase transition or quantum-mechanical tunnelling is involved, and it turns out that the termination of inflation is achieved more easily than in the previous model.

Let us begin by writing down the Lagrangian energy of the inflaton field:

$$L(\phi) = T - V = R^3\left[\frac{\dot{\phi}^2}{2} - V(\phi)\right] \tag{8.13}$$

where ϕ is the amplitude of the field, which in natural units, $\hbar = c = 1$, has the dimensions of mass, as in the case of the Higgs field (see Section 3.11), and R is the expansion factor. This equation involves the difference of the kinetic and potential energies T and V of the field as in (3.1). The total energy density of the field is then

$$\rho_\phi = \frac{(T+V)}{R^3} = \frac{\dot{\phi}^2}{2} + V(\phi) \tag{8.14}$$

The Euler–Lagrange equation (3.1) for the system takes the form

$$\frac{\partial}{\partial t}\left(\frac{\partial L}{\partial\dot{\phi}}\right) - \frac{\partial L}{\partial\phi} = 0 \tag{8.15}$$

Applying this to (8.13) and dividing through by R^3 gives

$$\ddot{\phi} + 3H\dot{\phi} + \frac{\mathrm{d}V}{\mathrm{d}\phi} = 0 \tag{8.16}$$

This equation is like that for a ball rolling to and fro in a saucer, or that of a simple pendulum oscillating in a very dense gas, the middle term corresponding to friction losses, that is, to the reheating mechanism at the end of inflation. If, at the beginning of the inflation process, the kinetic energy of the field is small compared with the potential energy, $\ddot{\phi} \approx 0$ and $\dot{\phi}$ is small, so that $\phi \approx \phi_0$, a more or less constant value, and $V = V(\phi_0) \approx \rho_\phi$. In this so-called slow roll approximation, the Friedmann equation (5.11) takes the form (using units $\hbar = c = 1$ and the relation $G = 1/M_{\mathrm{PL}}^2$):

$$H^2 = \frac{8\pi \mathrm{G}\rho_\phi}{3} = \frac{8\pi V(\phi_0)}{\left[3M_{\mathrm{PL}}^2\right]} \tag{8.17}$$

so that the universe inflates exponentially with an almost constant expansion factor as in (8.9). In this scenario, the potential near the minimum is often taken to be of the simple quadratic form

$$V(\phi) = \frac{1}{2}m^2\phi^2 \tag{8.18}$$

As the inflation proceeds, ϕ starts by changing slowly as V 'rolls' gently down the dashed part of the curve in Fig. 8.3. With $\ddot{\phi} \approx 0$, (8.16) and (8.18) give

$$\ddot{\phi} = \frac{-m^2\phi}{3H} \tag{8.19}$$

and integrating we obtain

$$\phi = \phi_0 \exp\left(\frac{-m^2\Delta t}{3H}\right) \tag{8.20}$$

where $\Delta t = t_2 - t_1$ is the period of inflation as in (8.11). Clearly, ϕ should not fall too rapidly or the full expansion will not be obtained, and inserting the limit from (8.11) we therefore find from (8.20) that

$$60 < H\Delta t < \frac{3H^2}{m^2}$$

and hence from (8.17) and (8.18) the condition

$$\frac{m}{M_{\mathrm{PL}}} < \left(\frac{2\pi}{15}\right)^{1/2} \tag{8.21}$$

This condition, imposed to ensure that a large enough inflation factor is obtained, also shows that the energy density in the inflaton field is comfortably less than M_{PL}^4, at which level unknown quantum gravitational effects could become important. Eventually the system rolls into the potential well of the true vacuum and inflation ceases, and as explained above, the to-and-fro oscillations in the well correspond to the reheating phase.

There are many other models of inflation, including those which incorporate supergravity, but we do not discuss them here. None seems yet to have been totally successful in providing exactly the conditions required, but there seems little doubt that an inflationary scenario of some sort is an essential first stage in the early development of the universe. There are two strong predictions of such models. The first is that the closure parameter Ω_{tot} must be extremely close to unity, that is, the curvature parameter k must be near zero, and the universe on very large scales must be flat. The second is that only one particular patch of the early universe, out of the many, would have been in the correct state at the chosen time when the quantum fluctuation described in the next section took place; there must have been many other universes growing from other such patches. So, enormous as is our universe, inflation suggests that it is but a dot in the ocean, a tiny part of a much larger space domain.

8.6 Quantum fluctuations and inflation

It is believed that quantum fluctuations are at the heart of anisotropies in the early universe. In Chapter 3, we saw that quantum fluctuations in elementary particle physics, in the form, for example, of the creation and annihilation of virtual electron–positron pairs, were able to account for the anomalous magnetic moments of the electron and the muon, and that such fluctuations are a vital part of the very successful electroweak theory. In a static universe, such virtual processes could not result in production of real particles, since pair creation will always be followed by annihilation. However, in the inflationary scenario, the rapid expansion implies that any virtual particle–antiparticle pairs which are created would not be able to annihilate completely. Both creation and annihilation rates are the product of particle densities and interaction cross-sections. So a lower particle density at annihilation than for the previous process of creation would lead to a net creation of real particles (from the energy in the inflaton field). This is the mechanism assumed for particle (and antiparticle) creation in the early universe. Such quantum fluctuations are also involved, for example, in connection with Hawking radiation from black holes (see Section 10.12).

Quantum fluctuations of course arise in the first place as a result of the uncertainty relation. In a particular time interval Δt the energy of a system cannot be specified to an accuracy better than ΔE, where $\Delta t \cdot \Delta E \sim \hbar$. Fluctuations in the inflaton field amplitude ϕ can be thought of as due to the different times at which different 'bubble' universes complete inflation as in Fig. 8.3, via the relation

$$\Delta t = \frac{\Delta\phi}{\dot{\phi}} \tag{8.22}$$

When discussing fluctuations in the microwave background radiation in Section 8.13 below, the amplitude of the fluctuations at the horizon scale are important, and they are determined by the different amounts that the universes have expanded:

$$\frac{\Delta\rho}{\rho} = \delta_{\text{hor}} = H\Delta t \sim \frac{H^2}{\dot{\phi}} \tag{8.23}$$

where the Hubble time is $1/H$ and we have used the relation $\Delta\phi \sim H$ from the uncertainty principle (again in units $h/2\pi = c = 1$). Using equation (8.17) for H^2 and (8.19) for $\dot{\phi}$, we thus obtain for the estimated density fluctuation

$$\frac{\Delta\rho}{\rho} \sim \left(\frac{m}{M_{\text{PL}}}\right)\left(\frac{\phi}{M_{\text{PL}}}\right)^2 \tag{8.24}$$

We repeat that experimentally this quantity is of order 10^{-5}. Ideally of course it would be nice to *predict* the magnitude of the fluctuations from the inflation model, but at the present time this does not seem possible, since the number expected depends on the precise form assumed for the inflaton potential $V(\phi)$.

Perhaps, in the course of future studies of polarization of the CMB induced by gravitational waves accompanying inflation, as described in Sections 8.15 and 8.16, it may prove possible to perfect the inflation model. At present, however, no definite predictions seem to be possible regarding the level of quantum fluctuations. Even so, the idea that the material universe, extending now to the order of 10^{26} m, had its origins in a quantum fluctuation, which started off space/time as we know it as a microcosm of radius 10^{-27} m is quite grandiose and appealing.

8.7 The spectrum of primordial fluctuations

The quantum fluctuations referred to above are ‘zero-point’ oscillations in the cosmic fluid. As soon as inflation commences, however, at superluminal velocity, most of the fluid will move *outside the horizon scale* $1/H$. (We recall here that the horizon distance is of order ct where t is the time after the beginning of the expansion, and in units $c = 1$ is equal to the reciprocal of the expansion rate $1/H$). This means that there will no longer be communication between the crests and the troughs of the oscillations: the quantum fluctuations are therefore *‘frozen’ as classical density fluctuations at the super-horizon scale*. We also note from (8.24) that since no particular distance scale is specified for the fluctuations, the spectrum of fluctuations should follow a power law, which (unlike an exponential, for example) does not involve any absolute scale. These fluctuations in density correspond to perturbations in the metric of space-time associated with variations in the curvature parameter. As discussed below, there are different possible types of fluctuation; however, it is usually assumed that the perturbations are *adiabatic*, that is, that the density variations are the same in different components (baryons, photons, etc.).

We can see how the fluctuations depend on the index of the power law determining the balance between small and large scales, using an argument due to Barrow (1988). This is based on the idea that an exponential expansion is invariant under a time translation. No matter at what time one fixes the start of the exponential growth, the universe will look the same at every epoch. Thus the expansion rate H is constant, the density ρ is constant, the horizon distance $1/H$ is constant, and the universe is effectively in a *stationary state*. No time or place can then have significance over any other, with the result that the amplitude of the perturbations in the metric structure must be the same on all length scales as they enter the horizon—otherwise a change in the magnitude of the perturbation could be used to indicate a time sense. This metric (curvature of space) is determined by the gravitational potential Φ, and in the absence of time dependence, this will obey Poisson’s Equation of Newtonian gravity (see (2.20)):

$$\nabla^2\Phi = 4\pi G\rho \tag{8.25}$$

Assuming spherical symmetry, $\nabla^2\Phi = \left(1/r^2\right)\left[\partial\left(r^2\partial\Phi/\partial r\right)\partial r\right]$, and the solution is

$$\Phi(r) = \frac{2\pi G\rho r^2}{3} \tag{8.26}$$

On the scale of the horizon distance $r_{\mathrm{Hor}} = 1/H$, which is the only natural length in the problem, we have therefore

$$\Phi = \frac{2\pi G\rho}{3H^2}$$

while on some arbitrary scale $\lambda < 1/H$, fluctuations in the gravitational potential due to fluctuations $\Delta\rho$ in density will be

$$\Delta\Phi = 2\pi G\Delta\rho\frac{\lambda^2}{3} \tag{8.27}$$

Hence the fractional perturbation in the gravitational potential on the scale λ has the value

$$\frac{\Delta\Phi}{\Phi} = H^2\lambda^2\frac{\Delta\rho}{\rho} \tag{8.28}$$

As explained above, in a stationary state, $\Delta\Phi/\Phi$ must be some constant independent of the arbitrary distance scale λ. Since H is also approximately constant, it follows that the density fluctuation, as it comes inside the horizon (and specifically its root mean square value) must have a spectrum with the power law dependence on λ of

$$\left\langle \delta_\lambda^2 \right\rangle^{1/2} = \left(\frac{\Delta\rho}{\rho}\right)_{\text{rms}} \sim \frac{1}{\lambda^2} \tag{8.29}$$

known as the PHZ (Peebles–Harrison–Zeldovich) spectrum, typical of the inflationary scenario. In words, this spectrum gives the universe the same degree of 'wrinkliness' and the same amplitude for the perturbations on the horizon, independent of the epoch, as would be expected for a stationary state. For this reason the above spectrum is called *scale-invariant*.

Notice that the fluctuations predicted are actually smaller than the purely statistical fluctuations on the number N of particles contained in the volume λ^3, since according to (8.29), $\Delta N/N \propto N^{-2/3}$, while for a statistical fluctuation $\Delta N/N \propto N^{-1/2}$. This smoothing out of fluctuations on large scales is an example of a general rule that, as we shall see below, in an accelerating universe, perturbations tend to decay, while in a decelerating universe, perturbations tend to grow with time.

Rather than describing the dependence of the fluctuations on the length scale λ, it is usual to discuss their Fourier decomposition in terms of the wavenumber $k = 2\pi/\lambda$. First we define the *density contrast* at space coordinate x, as the fractional deviation from the value $\langle\rho\rangle$ of the density averaged over the normalization volume:

$$\delta(x) = \frac{[\rho(x) - \langle\rho\rangle]}{\langle\rho\rangle}$$

The two-point correlation function for points in space separated by distance r is given by

$$\xi(r) = \langle\delta(x+r)\delta(x)\rangle$$

again averaged over all pairs of points in the volume. This correlation function is then expressed as a Fourier integral over the wave number k. Assuming that the phases of the fluctuations are random, cross terms will cancel and one obtains:

$$\xi(r) = \int |\delta(k)|^2 \exp(i\mathbf{k}\cdot\mathbf{r})\ \mathrm{d}^3k \tag{8.30}$$

Here, for simplicity, some factors of 2π in the definition of the Fourier transform have been omitted. The quantity $P(k) = |\delta(k)|^2$ is referred to as the *power* of the fluctuation spectrum. Assuming isotropy, one can integrate over polar and azimuthal angles:

$$\xi(r) = \iint |\delta(k)|^2 \exp(ikr\cos\theta)\ 2\pi\ \mathrm{d}(\cos\theta)k^2\,\mathrm{d}k$$

Taking the limits +1 and -1 for cos θ, the angular integration gives

$$\xi(r) = 4\pi \int P(k) \left[\frac{(\sin kr)}{kr}\right] k^2\ \mathrm{d}k \tag{8.31}$$

From this expression, we see that for values of $kr \gg 1$, the term in square brackets, and hence the integrand, will average to zero, while for values of $kr < 1$, the integral will vary as $k^3P(k)$. Because of the absence of any absolute scale, $P(k) \sim k^n$ must be represented by a power law, and for the inflationary model, we already know from the above analysis that $n = 1$. Thus $\xi \sim k^4$, and the square root of the correlation function then varies as k^2 or $1/\lambda^2$, as we already deduced in (8.29). As described below, analysis of the angular fluctuations of the CMB leads to an experimental determination of the index n from the WMAP experiment (see data summary by Yao *et al.* 2006):

$$n = 0.95 \pm 0.02 \tag{8.32}$$

and

$$\frac{\mathrm{d}n}{\mathrm{d}(\ln k)} = -0.003 \pm 0.010$$

This result is close to $n = 1$ as predicted by inflation. However, there are small deviations or 'tilt' $(1 - n)$ from this unit value, which are significant here at the 2σ level. These arise if one takes into account that during inflation $\ddot{\varphi}$ and $\dot{\varphi}$ are slightly different from zero. The result in the second line above indicates no significant 'running' of the index with the wave number.

8.8 Large-scale structures: gravitational collapse and the Jeans mass

In Chapter 5 the early universe was described as a homogeneous, isotropic and perfect primordial fluid (a perfect fluid being one in which frictional effects are negligible), undergoing a universal expansion. In contrast, the universe today is 'grainy' with the matter clumped into billions of individual galaxies, each containing of order 10^{11} stars, and separated from their neighbours by enormous voids in space. Starting off from the Big Bang, we have to ask what were the physical processes taking place which led to such structures. The developments on the smallest scales, that of the stars themselves, are dealt with in Chapter 10. Here we discuss the large-scale structures which were, we believe, originally seeded by tiny fluctuations in the inflationary phase described above, which are detectable today in observations on the microwave background radiation, as described in Sections 8.13–8.15. However, before discussing those observations and their interpretation, we consider the general conditions necessary for gravitational collapse of a classical gas cloud, as originally enunciated by Jeans.

Let us first estimate the time required for a cloud of ordinary gas to collapse under gravity, assuming to begin with that gas pressure can be neglected. Suppose that the cloud is spherical, of constant mass M and of initial radius r_0, and that it begins gravitational contraction. When the radius has shrunk to r, a small mass m in the outermost shell will have lost gravitational potential energy $GMm\,(1/r - 1/r_0)$ and gained kinetic energy $(m/2)\,(\mathrm{d}r/\mathrm{d}t)^2$, assuming that it was initially at rest. Equating these two, we get for the time of free fall from $r = r_0$ to $r = 0$

$$t_{\mathrm{FF}} = \int \frac{\mathrm{d}r}{\left(\mathrm{d}r/\mathrm{d}t\right)} = \int \left(\frac{2\mathrm{G}M}{r} - \frac{2\mathrm{G}M}{r_0}\right)^{-1/2} \mathrm{d}r \tag{8.33}$$

Substituting $r = r_0 \sin^2\theta$ and with the limits $\theta = \pi/2$ and 0, this integral gives

$$t_{\mathrm{FF}} = \left(\frac{\pi}{2}\right)\left(\frac{r_0^3}{2\mathrm{G}M}\right)^{1/2} = \left(\frac{3\pi}{32\mathrm{G}\rho}\right)^{1/2} \tag{8.34}$$

where ρ is the mean initial density of the cloud. Note that the result is independent of the radius, for a given initial density. This free fall time, it may be observed, is comparable with the circulation time of a satellite in close orbit about the initial cloud, equal to $(3\pi/\mathrm{G}\rho)^{1/2}$.

As the cloud of gas condenses, gravitational potential energy will be transformed into kinetic (heat) energy of the gas particles. If these are atoms or molecules, this motional kinetic energy may be absorbed through collisional dissociation of molecules or ionization of atoms, as well as resulting in atomic excitation which can be radiated away as photons if the cloud is transparent. These processes absorb and then re-emit the gravitational energy liberated and allow the cloud to contract further, but eventually hydrostatic equilibrium will be attained when the pressure of the heated gas balances the inward gravitational pressure. The total kinetic energy of the gas at temperature T will be

$$E_{\mathrm{kin}} = \left(\frac{3}{2}\right)\frac{MkT}{m} \tag{8.35}$$

where m is the mass per particle, M/m is the total number of particles, and $3kT/2$ is the mean energy per particle at temperature T. The gravitational potential energy of a sphere of mass M and radius r is

$$E_{\mathrm{grav}} \approx \frac{\mathrm{G}M^2}{r} \tag{8.36}$$

where there is a numerical coefficient of order unity, depending on the variation of density with radius (and equal to 3/5 if the density is constant). Comparing these two expressions, we find that a cloud will condense if $E_{\mathrm{grav}} \gg E_{\mathrm{kin}}$, that is, if r and ρ exceed the critical values

$$r_{\mathrm{crit}} = \frac{2M\mathrm{G}m}{(3kT)} = \left(\frac{3}{2}\right)\left(\frac{kT}{2\pi\rho\mathrm{G}m}\right)^{1/2} \tag{8.37}$$

$$\rho_{\mathrm{crit}} = \left(\frac{3}{4\pi M^2}\right)\left(\frac{3kT}{2m\mathrm{G}}\right)^3$$

Example 8.2 *Calculate the critical density and radius of a cloud of molecular hydrogen with a mass of 10,000 solar masses at a temperature of 20K.*

Inserting the values in SI units of $M_{\text{sun}} = 2 \times 10^{30}$, $\text{G} = 6.67 \times 10^{-11}$, $k = 1.38 \times 10^{-23}$ and $m = 3 \times 10^{-27}$ kg into equations (8.37) gives the values

$$r_{\text{crit}} = 1.07 \times 10^{19}\,\text{m} = 0.35\ \text{kpc}$$

$$\rho_{\text{crit}} = 3.83 \times 10^{-24}\,\text{kg}\,\text{m}^{-3} = 1150\ \text{mols}\,\text{m}^{-3}$$

These are typical temperature and density values for clouds of gas in *globular clusters*, which each contain of order 10^5 stars. Individual stars will form as a result of density fluctuations in the cloud, requiring from (8.37) gas densities about 10^8 times larger.

From the viewpoint of the development of large-scale structure in the universe, we want to determine which criteria lead to a cloud of gas condensing as a result of an upward fluctuation in density in one part of it. In terms of the density ρ, there is a critical size of the cloud called the *Jeans length* with a value

$$\lambda_{\text{J}} = v_{\text{s}} \left(\frac{\pi}{\text{G}\rho} \right)^{1/2} \tag{8.38}$$

obtained essentially by multiplying the sound velocity by the free-fall time. The mass of a cloud of diameter equal to the Jeans length is called the *Jeans mass*

$$M_{\text{J}} = \frac{\pi \rho \lambda_{\text{J}}^3}{6} \tag{8.39}$$

Here v_{s} is the velocity of sound in the gas. What do these equations mean? The typical time for sound waves (propagated as a result of any density perturbations) to cross a cloud of size L is L/v_{s}, and this is less than the gravitational collapse time (8.34) when $L \ll \lambda_{\text{J}}$. So the perturbation just results in sound waves oscillating to and fro, and there is no preferred location towards which matter can gravitate. On the other hand, if $L \gg \lambda_{\text{J}}$, sound waves cannot travel fast enough to respond to density perturbations and the cloud will start to condense around them. For a cloud of non-relativistic matter, the Jeans length λ_{J} in (8.38) and r_{crit} in (8.37) are of course one and the same (up to numerical factors of order unity). For then

$$v_{\text{s}}^2 = \frac{\partial P}{\partial \rho} = \frac{\gamma k T}{m} \tag{8.40}$$

where γ is the ratio of specific heats, equal to $5/3$ in neutral hydrogen. In that case

$$\lambda_{\text{J}} = \left(\frac{5\pi k T}{3\text{G}\rho m} \right)^{1/2} \tag{8.41}$$

Thus in terms of the temperature of non-relativistic gas particles, both r_{crit} in (8.37) and λ_{J} in (8.38) are of the order of magnitude $(kT/\text{G}\rho m)^{1/2}$.

8.9 The growth of structure in an expanding universe

We now apply the ideas in the preceding section, based on classical density perturbations, to fluctuations in the early universe. Suppose that an upward fluctuation in density occurs at some point in a static (i.e. non-expanding) homogeneous and isotropic fluid of non-relativistic particles, that is, the density increases by a small amount $\Delta\rho$ above the unperturbed density ρ, where $\Delta\rho \ll \rho$. The gravitational force which the perturbation exerts, and consequently the inflow of material attracted towards the perturbation per unit time, will both be proportional to $\Delta\rho$, so that $\mathrm{d}\,(\Delta\rho)/\mathrm{d}t \propto \Delta\rho$. This simple argument suggests that the density perturbation might be expected to grow exponentially with time. However, in the case of a non-static, expanding universe, the gravitational inflow can be counterbalanced by the outward Hubble flow. It then turns out that the time dependence of growth of the density fluctuation is a power law rather than an exponential. Intuitively, one can guess that, if the perturbation is small so that all effects are linear, and it is expressed in terms of the so-called *density contrast* $\delta = \Delta\rho/\rho$, this dimensionless quantity can only be proportional to the other dimensionless number associated with the Hubble flow, namely the expansion parameter ratio $R\,(t_2)/R(t_1)$ corresponding to times t_2 and t_1.

Quantitatively, we have to enquire whether the growth of cosmic structures on the largest scales can be understood in terms of the tiny anisotropies (temperature and density fluctuations at the 10^{-5} level) observed in the cosmic microwave radiation, already mentioned in Section 5.9 and discussed in more detail below. The standard treatment of the growth of small fluctuations in density by means of perturbation theory is rather lengthy and is given in Appendix C. Here we derive the principal result by means of a short cut, treating the initial upward density fluctuation as a matter dominated, closed 'micro-universe' of mass M and positive curvature k/R^2 with $k = +1$, as described by the lower curve in Fig. 5.4 and by Example 5.2. Then from equation (5.17) the values of R and t in parametric form are

$$R = a(1 - \cos\theta) = \left(\frac{a\theta^2}{2}\right)\left[1 - \frac{\theta^2}{12} + \cdots\right]$$

$$t = b\,(\theta - \sin\theta) = \left(\frac{b\theta^3}{6}\right)\left[1 - \frac{\theta^2}{20} + \cdots\right] \tag{8.42}$$

where $a = \mathrm{G}M/c^2$ and $b = \mathrm{G}M/c^3$, and the expansion on the right is for very early times, that is, for $\theta \ll 1$. Taking the 2/3 power of the second equation to find θ^2 as a function of $t^{2/3}$, and inserting in the first equation one obtains

$$R(t) = \left(\frac{a}{2}\right)\left(\frac{6t}{b}\right)^{2/3}\left[\frac{1 - \left[(6t/b)^{2/3}\right]}{20} + \cdots\right] \tag{8.43}$$

We see that when $t \ll b/6$, $R(t) \propto t^{2/3}$, that is, the increase in radius with time is the same as that in a flat, matter-dominated universe of $\Omega = 1$ (see Table 5.2).

For larger but still small values of t, the density enhancement, compared with the flat case, grows linearly with the expansion factor $R(t)$:

$$\delta = \frac{\Delta\rho}{\rho} = -\frac{3\Delta R}{R} = +\left(\frac{3}{20}\right)\left[\frac{6t}{b}\right]^{2/3} \approx \left(\frac{3}{10a}\right) R(t) \propto (1+z) \quad (8.44)$$

just as we anticipated. Incidentally, we may note here that, had we done the same exercise for an open universe as in (5.18), the value of δ would have come in with the opposite sign, with the density perturbation decreasing with time.

According to the simple linear dependence in (8.44), the primordial (10^{-5}) fluctuations in the microwave radiation at the time of decoupling ($z_{dec} \sim 1000$) would by now have grown by some three orders of magnitude in a matter-dominated universe. This, however, is not enough to account for the much larger density fluctuations in the material of the present universe. The conclusion is that the observed level of fluctuations in the microwave radiation would have been too small to account for the observed structures in terms of growth of fluctuations in the baryonic component alone and that non-relativistic (cold) dark matter is also needed.

8.10 Evolution of fluctuations during the radiation era

So far, we have been considering the growth of fluctuations in non-relativistic matter, that is, the baryonic component and so-called cold dark matter. However, before the time of decoupling at $z \sim 1100$, the energy density would have contained contributions from the 'relic' primordial photons and neutrinos/antineutrinos discussed in Sections 5.10–5.12. These would indeed have made dominant contributions at the earlier stages of the radiation era, before the time of matter–radiation equality (i.e. for $z > 3000$).

Figure 8.4 shows an early plot of the values of the root mean square amplitude of density fluctuations plotted against the distance scale λ introduced in (8.29). On very large scales (typically angular ranges of 10°–100°) these were deduced from the temperature fluctuations of the microwave background, as first observed by the COBE satellite experiments (Smoot *et al.* 1990), which would of course reflect the density fluctuations in matter (see Fig. 8.8).

On smaller scales and angles, the density fluctuations have been deduced from analysis of large-scale galaxy surveys. At large scales, the spectrum of fluctuations does seem to follow quite well the $1/\lambda^2$ variation predicted by inflation in (8.29), the universe becoming progressively smoother over the largest distances, while the spectrum flattens off at the smaller scales of galaxies and galaxy clusters, that is, for $\lambda < 100$ Mpc. The curves show the expected amplitude for a cold dark matter scenario, and one where hot and cold dark matter are mixed. The flattening of the curves at small scales is due to the damping effects of the relativistic particles—photons and neutrinos—which we now discuss.

In the early stages of the Big Bang, the universe was radiation-dominated and the velocity of sound was relativistic, with a value $v_s \sim c/\sqrt{3}$—see Table 5.2. This means that, using (5.47), with $\rho_r c^2 = \left(3c^2/32\pi G\right)/t^2$, the Jeans length

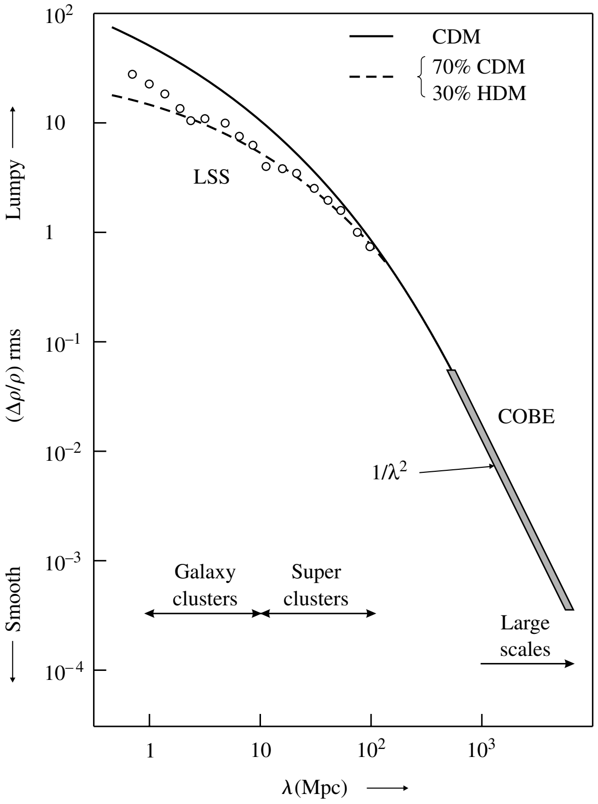


Fig. 8.4 A plot of density fluctuations against the scale λ, from the COBE satellite observations on the microwave background at large angular scales, and from galaxy surveys of large-scale structure (e.g. the infrared survey from the IRAS satellite experiments) on smaller angular scales. Essentially, such surveys consist of counting the number of galaxies contained in each of many volumes λ^3 of sky, determining the r.m.s. fractional fluctuation about the average number, and repeating the process for different values of λ. The curves show the early predictions from cold and mixed dark matter models (after Kolb 1998).

was

$$\lambda_J = c\left[\frac{\pi}{(3G\rho_r)}\right]^{1/2} = ct\left(\frac{32\pi}{9}\right)^{1/2} \tag{8.45}$$

Thus the horizon distance and the Jeans length are both of order *ct* during the radiation era.

8.10.1 The photon component

We now discuss the development with time of the mass inside the horizon during the radiation-dominated era, and consider whether, during that period, initial density fluctuations in matter could have survived. First, we consider the role of the photon component. The actual baryonic mass inside the horizon during this era would be

$$M_H(t) \sim \rho_b(t)(ct)^3 \propto \frac{1}{T^3} \tag{8.46}$$

where the T dependence arises from the fact that $\rho_b \propto 1/R^3 \propto T^3$ and from (5.49), $t \propto 1/T^2$ during the radiation-dominated era. This dependence will of course flatten off near the decoupling temperature because of the increasing effect of the baryons in reducing the sound velocity. At the time of baryon–photon decoupling, $z_{\rm dec} \sim 1100$, $\rho_b{=}\rho_c\Omega_b(1+z_{\rm dec})^3$, and

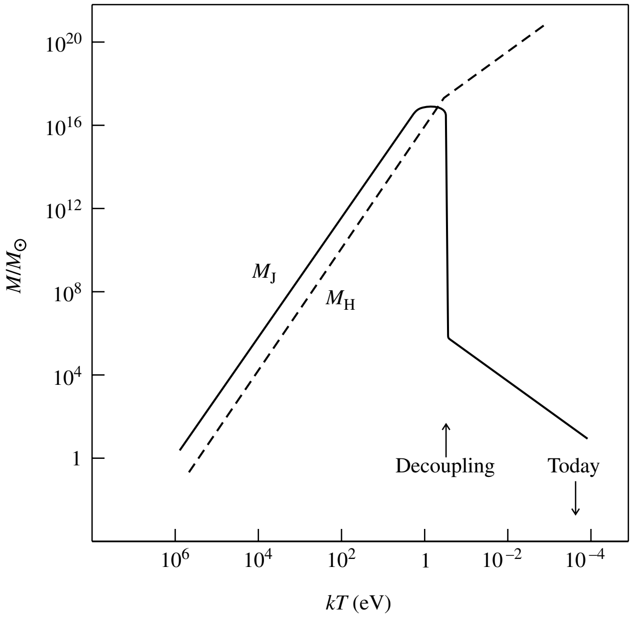


Fig. 8.5 The variation with radiation temperature T of the (baryonic) mass M_H inside the horizon (i.e. inside the largest distance over which causal effects are possible) and of the Jeans mass M_J (i.e. the smallest mass which can overcome the pressure of radiation and contract under gravity). After decoupling of matter and radiation, the Jeans mass falls abruptly as the velocity of sound reduces by a factor of 10^4, while the mass inside the horizon continues to increase (as $1/T^{1\dots5}$).

$t_{\text{dec}} = t_0/(1 + z_{\text{dec}})^{3/2} \sim 10^{13}$ s. Inserting the value of ρ_c from (5.26) we find

$$M_H\,(t_{\text{dec}}) \sim 10^{18}\Omega_b M_{\text{sun}} \sim 10^{17} M_{\text{sun}} \tag{8.47}$$

for $\Omega_b = 0.04$. The Jeans mass will be an order of magnitude larger. This demonstrates that fluctuations on the scale of galaxies ($M \sim 10^{11} M_{\text{sun}}$) and clusters ($M \sim 10^{14} M_{\text{sun}}$) come inside the horizon during the radiation era, at redshifts $(1 + z) \sim 10^5$ and $\sim$2000 respectively. The variation of M_H and M_J with T is shown in Fig. 8.5.

As stated above, there are in principal several different possible types of fluctuation. *Adiabatic* fluctuations behave like sound waves, with baryon and photon densities fluctuating together, while for *isothermal* fluctuations the matter density fluctuates but the photon density does not, so that matter is in a constant temperature photon bath. Or it could be that both matter and photon densities fluctuate but with opposite phases in *isocurvature* fluctuations. Present indications are that adiabatic fluctuations are most likely to be dominant. What happens depends on the scale of distance considered. While the matter is non-relativistic, photons travel at light velocity and through radiation pressure, which opposes gravitational infall, can stream away from regions of higher density to ones of lower density, so ironing out any fluctuations. During most of the radiation era, the photon energy density is larger than that of the baryons or of 'cold' dark matter, so that if the photons diffuse away, the amplitude of the fluctuation will be severely reduced, a process called *diffusion damping* or *Silk damping*.

This loss of photons will be prevented if they are locked into the baryonic matter by 'Thomson drag', that is, by Compton scattering by electrons of the baryon–electron plasma, which at the energies concerned is determined by the Thomson cross-section (1.26d).

It is found, as shown in Example 8.3 below, that fluctuations containing baryonic masses well above $10^{13} M_{\text{sun}}$—that is, the size of galaxies or larger

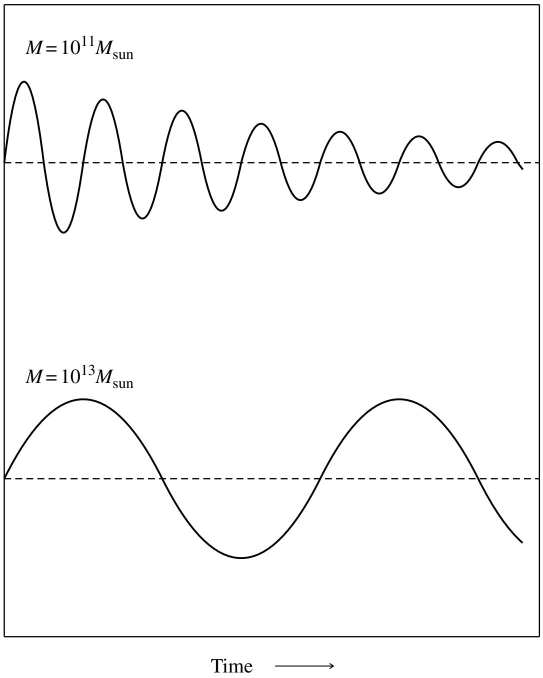


Fig. 8.6 During the radiation era, adiabatic fluctuations of wavelength encompassing baryonic masses below $10^{12}M_{\text{sun}}$ are damped out by the leakage of the photon component, while those of masses above $10^{14}M_{\text{sun}}$ remain at a practically constant amplitude until the epoch of matter–radiation decoupling (see Example 8.3).

objects—will survive without significant reduction of amplitude to the era of decoupling, after which they can grow. Fluctuations on smaller scales will, on the contrary, be ironed out, as indicated in Fig. 8.6.

Example 8.3 *Estimate the minimum mass associated with a primordial density fluctuation which could survive to the era of decoupling, taking only account of photon damping.*

The scattering mean free path of photons through ionized baryonic matter will be $l = 1/(n_e\sigma)$, where the electron number density $n_e \sim \rho_b N_0$, N_0 is Avogadro's number and σ is the Compton cross-section for $\gamma e \to \gamma e$, equal at these energies to the Thomson cross-section (1.26d). Since the scattering is isotropic, the result of N successive scatters is that the photon travels a bee-line distance D where $\langle D^2 \rangle = (l_1 + l_2 + l_3 + \cdots + l_N)^2 = N\langle l^2 \rangle$, since the cross-terms cancel in the square (this is an example of the famous 'drunkard's walk' problem).

The time taken for the photon to cover a bee-line distance D is therefore $t = Nl/c$, so that $D = (ct \cdot l)^{1/2}$ is the geometric mean of the horizon distance and the scattering mean free path. Hence the time required for a photon to diffuse out of a fluctuation of scale length D is of order D^2/lc, and if this is much less than the time t since the onset of the Big Bang, the photons will stream away and the fluctuation will be damped out. In order for the fluctuation to survive until $t = t_{\text{dec}}$ (and thereafter grow), we therefore need

$$D^2 > lct_{\text{dec}} = \frac{ct_{\text{dec}}}{(\rho_b N_0 \sigma)} \tag{8.48}$$

For a lower limit to D we take ρ_b at $t = t_{\text{dec}}$, equal to $\rho_b(0)(1 + z_{\text{dec}})^3$, giving $M > D^3\rho_b \sim 10^{13}M_{\text{Sun}}$. The corresponding scale length today is $D(1 + z_{\text{dec}}) \sim 10$ Mpc.

8.10.2 The neutrino component

In addition to the photon relics of the Big Bang, there are also relic relativistic neutrinos—'hot dark matter'—in comparable numbers, as discussed in Section 5.10, and they will also have a crucial effect on the development of structure. Neutrinos are not constrained by Thomson scattering off electrons in the plasma, since their weak collision cross-section at the energies concerned, with kT of order of a few eV, is only about 10^{-56} cm^2 for $\nu p \rightarrow \nu p$ scattering (i.e. for neutral current scattering, since they are below the energy threshold for charged current interactions) and is even less for $\nu e \rightarrow \nu e$ scattering. As stated in Section 5.11, relic neutrinos would have had essentially no further interaction with matter after they have cooled by expansion to $kT \sim 3$ MeV, only 0.1 s after the Big Bang.

Like any other dark matter, relativistic neutrinos will cluster gravitationally if they are able to do so. But, if they have sufficient time they can also stream freely away from local upward fluctuations in density, in *collisionless damping*. Since they have velocities close to c, they can stream up to distances equal to the optical horizon at that time. In fact what matters here is the distance of the horizon at the time t_{eq} of equal energy density of radiation and matter, since perturbations coming inside the horizon during the time of radiation domination will not be able to grow because of neutrino damping effects. From the formula for horizon distance in Chapter 5, one finds that this is of order 150 Mpc. For such scales, neutrinos will tend to damp down any upward fluctuations in density, so that such fluctuations will fail to grow. For fluctuations significantly above this size, say 400–500 Mpc, neutrinos will cluster like other dark matter, since they cannot escape the distances covered by the fluctuation in the time available.

In Figs. 8.4 and 8.7 we note that indeed such 'hot dark matter' hardly affects the fluctuation spectrum on distance scales exceeding about 400 Mpc. Such scales enclose masses of order $10^{16}M_{\text{sun}}$, that is, the size of superclusters. Thus in a neutrino-dominated early universe, superclusters would be the first to form. From the early results shown in Fig. 8.4, taken from the first edition of this book, the contribution of 'hot' dark matter to the total dark matter was estimated to be of order 30%, so that, since the value of kT towards the end of the radiation era was of order of a few eV, the fact that the neutrinos were 'hot', that is, relativistic, suggests an upper limit to neutrino masses of order 1 eV/c^2 or so. In the next section we quote more precise limits from more recent experiments.

8.11 Cosmological limits on neutrino mass from fluctuation spectrum

The observed spectrum of fluctuations we have described above has indeed led to important limits on the neutrino mass, if the results of different types of experiment are combined. Qualitatively one can understand the effects

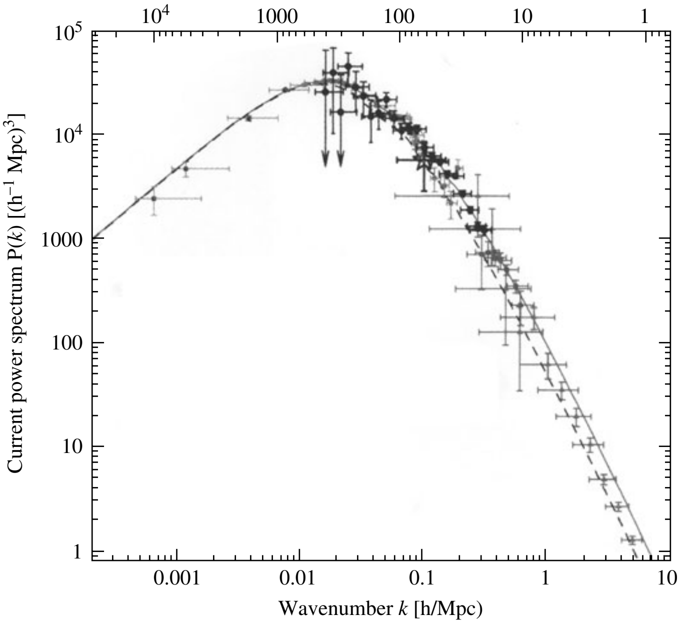


Fig. 8.7 Plot of the power $P(k)$ against the wavenumber k (quoted in reciprocal Mpc) from the CMB COBE survey at small k (large distances $\lambda > 1000$ Mpc) and from galaxy surveys at larger k values. These include the Sloane Digital Sky Survey of some 250,000 galaxies, weak lensing, and Lyman alpha forest data. The scales involve the parameter $h = 0.72$, which is the current Hubble constant H_0 divided by 100. Thus the upper scale is of 0.72 λ, where λ is the scale length in Mpc. The full line is the theoretical prediction for cold dark matter and zero neutrino mass, the dashed line for a neutrino mass of 1 eV/c^2 (from Tegmark 2005).

as follows. First, as pointed out above, and irrespective of neutrino mass, fluctuations on scales above 400 Mpc will be virtually unaffected, and the power $P(k) \sim k^1$ as in (8.29). On smaller and smaller scales, neutrino streaming has an ever bigger effect in ironing out the fluctuations, with the result that $P(k)$ decreases as k increases. As shown in Fig. 8.7, $P(k) \sim 1/k^2$ for $k \sim 1$ Mpc^{-1}, that is to say the mean square fluctuation $k^3P(k)$ is proportional to k, and the root mean square fluctuation to $k^{1/2}$, or $1/\lambda^{1/2}$.

However, although it has little effect on the shape of the spectrum, the actual level of the fluctuation at small scales does depend quite critically on neutrino mass. As indicated above, for masses of order 1 eV/c^2, the neutrino mass energy and thermal kinetic energy are comparable. Let us recall here that, as indicated in Chapter 5, the redshift at decoupling of photons and baryons is $z_{\text{dec}} = 1100$, when $kT = 0.3$ eV, so that at the time of matter–radiation equality, when $z_{\text{eq}} \sim 3000$, $kT \sim (3000/1100) \times 0.3 \sim 1$ eV. For masses much larger than 1 eV/c^2, the neutrino velocity could therefore be significantly less than c and the streaming distance correspondingly reduced, but more importantly, the contribution Ω_ν of neutrinos to the overall density parameter is proportional to neutrino mass and larger masses therefore have a bigger proportionate effect in damping out fluctuations. Roughly, an increase in neutrino mass from zero to 1 eV/c^2 is found to reduce the fluctuation level by about a factor 2—see Fig. 8.7.

The recent analyses of neutrino mass effects take into account the details of several large-scale galaxy surveys, for example, the Sloane Digital Sky Survey (SDSS) of Doroshkevich *et al.* (2003) of over 250,000 galaxies using the Hubble Space Telescope, and the 2 Degree Field Galaxy Redshift Survey (2dFGRS) of Elgaroy *et al.* (2002). Also included are results from 'Lyman alpha forest' studies (see Section 9.14) at the smallest scales, as well as the cosmic parameters emerging from the WMAP and other analyses of 'acoustic peaks' in the angular spectrum of the microwave radiation, as discussed later in this chapter. The important point about this combined information of different cosmological parameters, is that it provides the normalization between small scale fluctuations from galaxy surveys, and the large-scale fluctuations from the COBE experiments on the microwave background, thus placing stronger constraints on the neutrino contribution. Recent analyses (Tegmark 2005) are consistent with zero neutrino mass and lead to an upper bound for the mass summed over all flavours (90% confidence level):

$$\sum m_\nu(e, \mu, \tau) < 0.42 \ \frac{\mathrm{eV}}{\mathrm{c}^2} \tag{8.49}$$

This limit is already almost within an order of magnitude of the minimum mass values from neutrino oscillations (see Chapter 4), and will surely improve in the near future.

Incidentally, at this point it may be remarked that *today*, relic neutrinos have no chance of clustering around galaxies or clusters. They are non-relativistic, with a value of $kT = 0.17$ meV (see Section 5.10) compared with mass differences from (4.12)—and by inference the masses themselves—of up to at least 10–50 meV. The mean velocity of their Maxwell energy distribution would be $\sim$6000 km s^{-1} for a neutrino mass of 1 eV/c^2, and $\sim$20,000 km s^{-1} for a mass of 0.1 eV/c^2. Either of these velocities is way beyond the escape velocity from galaxies or clusters (see Problem 8.8).

8.12 Growth of fluctuations in the matter-dominated era

As indicated in Example 8.4 below, as soon as matter and radiation decoupled and neutral atoms formed, the velocity of sound, and hence the Jeans length, decreased by over 10,000 times. Growth of inhomogeneities on galactic and smaller scales then became possible.

Example 8.4 *Estimate the value of the Jeans mass just after the epoch of decoupling of matter (baryons) and radiation.*

After decoupling of baryonic matter from radiation and the recombination of protons and electrons to form atoms, the velocity of sound is given by (8.40):

$$\frac{v_\mathrm{s}^2}{c^2} = \frac{5kT}{3m_\mathrm{H}c^2}$$

Taking $kT = 0.3$ eV from Section 5.12 and the mass energy of the hydrogen atom $m_\mathrm{H}c^2 = 0.94$ GeV, we find

$$\frac{v_\mathrm{s}}{c} = 2 \times 10^{-5}$$

Thus, compared with the pre-decoupling epoch, the sound velocity and the Jeans length have fallen by a factor of 10^4, and the Jeans mass from $10^{18} M_\mathrm{sun}$ to $10^6 M_\mathrm{sun}$. This last is the typical mass of *globular clusters* of order 10^5 stars, which are some of the oldest objects in the sky (see Chapter 10). Obviously, even larger objects such as galaxies and galaxy clusters would have had no problem in condensing under gravity at this epoch.

We can see from the above discussion that only after the universe became matter dominated did perturbations really have a chance to grow. In this respect, dark matter, and specifically cold dark matter, plays a vital role in the development of structures at the galactic and super-galactic level. It turns out in fact that the calculated increase in density contrast with time, as described above and starting out from values of $\Delta\rho/\rho \sim 10^{-5}$ following inflation, is not enough to account for the observed growth of galaxies and clusters, relying simply on the gravitational collapse of the baryonic component alone (with $\Omega_b(0) \sim 0.04$), once it has decoupled from radiation at a redshift $z \sim 1100$. As previously stated, one requires substantial amounts of cold (i.e. non-relativistic) dark matter (with $\Omega_{\rm dm}(0) \sim 0.20$) as a component of the primordial universe. Unlike normal (baryonic) matter, this will not interact with radiation *via* Thomson scattering and it also begins to dominate over radiation at an earlier epoch ($z \sim 3000$–5000, see Section 5.13), and thus is both more efficient at achieving gravitational collapse and has more time to achieve it. Of course, once the dark matter agglomerations have formed gravitational potential wells, baryons will fall into them and indeed their increase in density contrast will follow that of the dark matter.

Finally, we note that the fluctuation spectrum, for example, that in Fig. 8.4, does seem to follow very well the $1/\lambda^2$ variation predicted on very large scales by the inflation scenario in (8.29), the universe becoming progressively smoother over the largest distances, while the spectrum flattens off at the smaller scales of galaxy clusters. This dependence on λ corresponds to the power law $P(k) \sim k^n$ in k-space, with $n \approx 1$ as in (8.32), providing strong verification of the inflation scenario.

8.13 Temperature fluctuations and anisotropies in the CMB

So far we have been discussing the dependence of the density fluctuations of matter on scale length λ or wave number k. However, much of our present knowledge of the basic parameters of the universe comes from the detailed and increasingly precise studies over the last 10 or 15 years of tiny (10^{-5}) *temperature variations* in the CMB in different directions in space, that is, their angular dependence. Of course, via their interactions with the baryons, the photons will suffer density and therefore temperature fluctuations which will mirror the fluctuations in matter density. We first briefly discuss the sources of these anisotropies, followed by a description of the experimental situation and its analysis.

The largest (10^{-3}) effect observed is from the dipole term, mentioned in Chapter 5, with the Doppler shift in temperature $T(\theta) = T_0\,[1 + (v/c)\cos\theta]$, arising from the 'peculiar velocity' vector $\mathbf{v}$ ($= 370\,\mathrm{km\,s^{-1}}$) of the Earth relative to the Hubble flow. The COBE satellite observations also detected higher multipoles, which corresponded to variations in the temperature over angular scales of 7° or more, the scale being set by the detector resolution. Figure 8.8 shows all-sky maps of the observed temperature fluctuations in the CMB, from the COBE and the more recent WMAP satellite experiments, the latter with angular resolution of order of a few minutes of arc. The temperature fluctuations

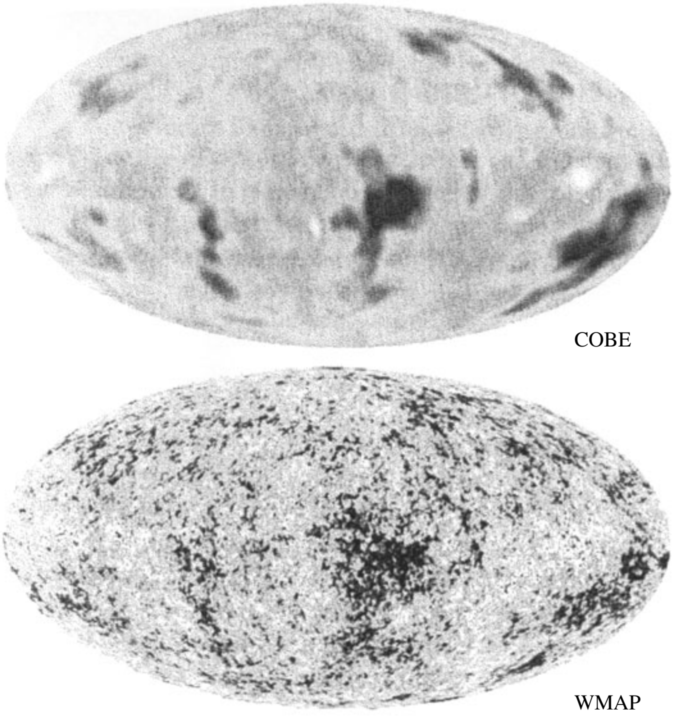


Fig. 8.8 All-sky maps of the temperature fluctuations in the CMB as measured by (a) the COBE experiments (Smoot *et al.* 1990), and (b) the WMAP experiment (Bennett *et al.* 2003). The increased resolution in WMAP is apparent. In these displays, the contrast has been enhanced about 0.3 million times, to make the tiny (10^{-5}) fluctuations visible.

observed in these higher multipoles were tiny, of order 10^{-5}, with an r.m.s. value of 18 μK. There are in fact two types of anisotropy: *primary anisotropy*, due to effects occurring before or during the last act of scattering of the microwave radiation before decoupling at $z_{\rm dec} = 1100$; and *secondary anisotropy,* due to interactions of the photons with interstellar gas or gravitational potentials, on their 14 billion year journey through space to present-day detectors. First, these will tend to smear out the primary anisotropies at the smallest angles. Secondly, because of the primary anisotropies, Thomson scattering off free electrons in any intergalactic plasma will produce plane polarization, particularly at large angles of scattering. This large angle polarization has been observed and gives important limits on the re-ionization of the interstellar medium after formation of the first stars, as described below.

8.13.1 The primary anisotropies

The primary anisotropies (variations in temperature with direction in space) in the CMB result from the gravitational (and to a lesser extent, electromagnetic) interactions of the photons with matter, occurring in the final acts of scattering before radiation and matter decoupled (at $z_{\rm dec} \sim 1100$), when the universe was already matter dominated. There are several different effects, but as we shall see below, large scales are dominated by gravitational interactions in what is known as the *Sachs-Wolfe effect*. First, if there is an upward fluctuation in the density of matter, the photons in that region will be red-shifted (cooled) as they climb out of the gravitational potential well, with $\Delta T/T = \Delta\Phi/c^2$, where the change $\Delta\Phi$ in the gravitational potential is negative. On the other hand, this change in potential causes a time dilation effect, $\Delta t/t = -\Delta\Phi/c^2$ (recall that, in a gravitational field, clocks run slow). Since the scale parameter R varies as $t^{2/3}$ in this matter-dominated phase, and R also varies as $1/T$, then $\Delta T/T = -\,(2/3)\ \Delta t/t = -\,(2/3)\ \Delta\Phi/c^2$. This means the photons are heated because they come from a region corresponding to an earlier and hotter time. The net effect of these two terms is a Sachs-Wolfe (SW) cooling:

$$\left(\frac{\Delta T}{T}\right)_{\mathrm{SW}} = \frac{\Delta\Phi}{3c^2} \tag{8.50}$$

There are several other causes of primary anisotropies. Since the plasma is in motion, there can be Doppler smearing of the photon frequency and hence temperature. However, as indicated above, the most important other source appears to be the *adiabatic effect,* in which the photons and matter fluctuate together. In an over-dense region, the radiation therefore becomes compressed and hotter. Since the photon number fluctuation $\Delta n_\gamma/n_\gamma = \Delta\rho/\rho$ and $n_\gamma \sim T^3$, it follows that

$$\left(\frac{\Delta T}{T}\right)_{\mathrm{AD}} = \left(\frac{1}{3}\right)\frac{\Delta\rho}{\rho} \tag{8.51}$$

where $\rho = \rho_m$ is the matter density.

The ratio of these two effects, using the expression (8.29) relating the variation in the gravitational potential due to a variation in density on the scale λ, is

$$\frac{(\Delta T/T)_{\mathrm{SW}}}{(\Delta T/T)_{\mathrm{AD}}} = \frac{2\pi \mathrm{G}\rho\lambda^2}{3c^2} \tag{8.52}$$

so that the SW effect was greater at scales such that at the time of decoupling

$$\lambda^2(\mathrm{dec}) > \frac{3c^2}{\left[2\pi \mathrm{G}\rho_m(0)\,(1+z_{\mathrm{dec}})^3\right]}$$

or for the situation today, multiplying λ by the scale factor $(1 + z_{\mathrm{dec}})$ and expressing the relation in terms of the critical density (5.26) and the matter closure parameter:

$$\lambda > \frac{(2c/H_0)}{[\Omega_m(0)\,(1+z_{\mathrm{dec}})]^{1/2}} \approx 0.50\ \mathrm{Gpc} \tag{8.53}$$

for $\Omega_m(0) = 0.26$, $z_{\mathrm{dec}} = 1100$. Thus the SW anisotropies in CMB temperature dominate at distance separations above 1 Gpc or, taking the present optical horizon as 15 Gpc (see (5.45)), over angular separations above 3° or so.

8.13.2 Secondary anisotropies

Immediately after the decoupling of baryonic matter and photons (i.e. the 'recombination' of electrons with hydrogen ions) at $z_{\mathrm{dec}} = 1100$, the universe would have consisted of photons, neutrinos, and neutral gas atoms of hydrogen and helium, together of course with the ubiquitous dark matter and dark energy.

During the ensuing 'dark ages', matter would have increasingly clumped together gravitationally, until 'protostars' first formed, followed eventually by stellar fusion processes (thermonuclear reactions), as described in Chapter 10. The dark ages then came to an end and there was light. The ultraviolet light emitted by these first stars would have re-ionized the interstellar gas. It turns out that secondary anisotropies of the CMB can become important if re-ionization commences as a result of star formation at $z \sim 12$ or less (see Problem 8.7). The WMAP (Wilkinson Microwave Anisotropy Probe) results (Alvarez *et al.* 2006) indeed find experimental evidence for this. They measured the polarization arising from Thomson scattering of the CMB from free electrons of the ionized plasma at large angles. CMB polarization is discussed below (Section 8.16). What the above paper quoted was an optical depth (effectively the scattering probability via the Thomson cross-section) of 0.09 ± 0.03 and a re-ionization redshift of $z = 11$.

Apart from the polarization effects of Thomson scattering, the CMB photons, on their way to the observer, may happen to traverse patches of dense hot plasma, in which case they can be blue-shifted as a result of collisions with energetic electrons (the *Sunyaev-Zeldovich effect*). Furthermore, if the photons traverse regions of variable gravitational potential, they will suffer gravitational shifts of frequency (the so-called integrated SW effect). All these processes will result in the smearing out of the primary anisotropies, most notably at very small angles.

8.14 The angular spectrum of anisotropies: 'acoustic peaks' in the distribution

In the discussion of the horizon distance $D_{\rm H}$ in Section 8.3, we concluded that, for a flat universe ($k = 0$), the angle subtended today by the (optical) horizon at the time of decoupling of matter and radiation was of order one degree, as in (8.7). A pressure wave can arise from density inhomogeneities and the interplay of the gravitational attraction and compressional effects of non-relativistic matter, on the one hand, opposed by photon pressure on the other. The propagation of such a pressure wave depends on the velocity of sound $v_{\rm s}$, and the acoustic horizon is $v_{\rm s}/c$ times the optical horizon distance. If the cosmic fluid is radiation-dominated at this epoch, then from Table 5.2 this ratio is approximately $1/\sqrt{3}$. The acoustic horizon at the time of last scatter of the microwave radiation in this case subtends today an angle at the Earth of approximately

$$\theta_{\rm acoustic} \sim \frac{ct_{\rm dec}\,(1 + z_{\rm dec})}{\sqrt{3}c\,(t_0 - t_{\rm dec})} \sim 1^\circ \tag{8.54}$$

Coming now to observational matters, the fluctuations in the temperature of the microwave radiation are measured as a function of position in the sky, and the correlation determined between two points separated by a particular angle θ. Suppose a measurement of the radiation temperature in a direction specified by a unit vector $\mathbf{n}$, as compared with the whole-sky average T, indicates a deviation $\Delta T(\mathbf{n})$, while in a direction $\mathbf{m}$ it is $\Delta T(\mathbf{m})$. The correlation between pairs of points in the sky is given by the average quantity

$$\begin{aligned} C(\theta) &= \left\langle \left(\frac{\Delta T\,(\mathbf{n})}{T} \right) \left(\frac{\Delta T\,(\mathbf{m})}{T} \right) \right\rangle \qquad \text{with } \mathbf{n}\cdot\mathbf{m} = \cos\theta \\ &= \sum (2l+1)\, C_l\, \frac{P_l(\cos\theta)}{4\pi} \end{aligned} \tag{8.55}$$

where the average is taken over all pairs of points in the sky separated by angle θ. In the second line, the distribution $C(\theta)$ has been expanded as a sum of Legendre polynomials $P_l(\cos\,\theta)$ running over all values of the integer l. The coefficients C_l describe the fluctuation spectrum, which depends not only on the initial spectrum of density fluctuations as discussed in Section 8.7, but on several other parameters, such as the baryon–photon ratio, the amount of dark matter, the Hubble constant, and so on. As had been predicted over 30 years ago, the measurement of the values of C_l for l-values in the hundreds should determine these parameters (the coefficient C_l for $l = 1$ corresponds to the dipole term mentioned above and is disregarded here).

The Legendre polynomial $P_l(\cos\,\theta)$ in (8.55) oscillates as a function of θ between positive and negative values, having l zeros between 0 and π radians, with approximately equal spacing

$$\Delta\theta \approx \frac{\pi}{l} \approx \frac{200}{l} \text{ degrees} \tag{8.56}$$

The sum $\Sigma(2l+1)P_l(\cos\,\theta)$ from $l = 1$ to $l_{\text{max}} \gg 1$ has a strong maximum in the forward direction ($\theta = 0$), where the amplitudes of all the different l-values add, while at larger angles the various contributions largely cancel out, the amplitude falling off to practically zero within the angular interval $\Delta\theta = 200/l_{\text{max}}$ degrees. As (8.54) indicates, fluctuations over an angular range of a degree or less are relevant to the acoustic horizon distance at decoupling and will therefore be concerned with polynomial contributions of $l > 100$.

Suppose now that an initial 'primordial' perturbation in the density of the 'photon–baryon fluid' occurs. This can be decomposed into a superposition of modes of different wavenumber k and wavelength $\lambda = 2\pi/k$. If the wavelength becomes larger than the horizon size—as it certainly would in the course of inflation—then the amplitude of that mode of the perturbation will become *frozen*; there could no longer be a causal connection between the troughs and the crests. However, as time evolves, $\lambda(t)$ will increase with the scale parameter $R(t) \propto t^n$, where $n = 2/3$ for a matter-dominated situation and $n = 1/2$ for radiation domination. Hence $\lambda(t)/D_{\text{H}}(t) \sim 1/t^{(1-n)}$ as in (8.6) and, since $n < 1$, it follows that in time, λ will come inside the acoustic horizon and the amplitude of that mode will then start to oscillate as a *standing acoustic wave* in the cosmic fluid. Modes of smaller wavelength will enter the horizon earlier and oscillate more quickly (since the frequency varies as $1/\lambda$). The effect of having components of different wavelengths and phases is that one obtains a series of *acoustic peaks* when the amplitude is plotted as a function of l, as shown in Figs. 8.9 and 8.12, with the first peak corresponding to a wavelength equal to the horizon distance. The quantity $l(l+1)C_l$ is plotted against l (since it turns out that this is a constant, independent of l, for a scale-invariant spectrum of fluctuations).

Example 8.5 *Estimate the magnitudes of the physical objects corresponding to the 'acoustic peaks' in the angular power spectrum of the microwave radiation*

The amount of material contained inside the horizon at the epoch under consideration, that is, when matter decoupled from radiation and protons started to recombine with electrons to form atoms and molecules, would in fact be somewhat larger than the typical mass of a supercluster (Fig. 8.4). The baryonic mass inside the horizon (since $\rho_b \sim \rho_r$ at $t = t_{\text{dec}}$) is

$$M_{\text{hor}} \sim (ct_{\text{dec}})^3\ \rho_r(\text{dec}) \sim 10^{17}\, M_{\text{sun}} \tag{8.57}$$

using the fact that $z_{\text{dec}} \sim 1100$, $t_{\text{dec}} \sim 10^{13}$ s (see Section 5.13), and $\rho_r(\text{dec}) = \rho_r(0)(1 + z_{\text{dec}})^4 \sim 10^{-19}$ kg m^{-3}.

The first peak in Fig. 8.9, at $l \sim 200$, corresponds to the mode which has just come inside the horizon and compressed only once, the second corresponds to a

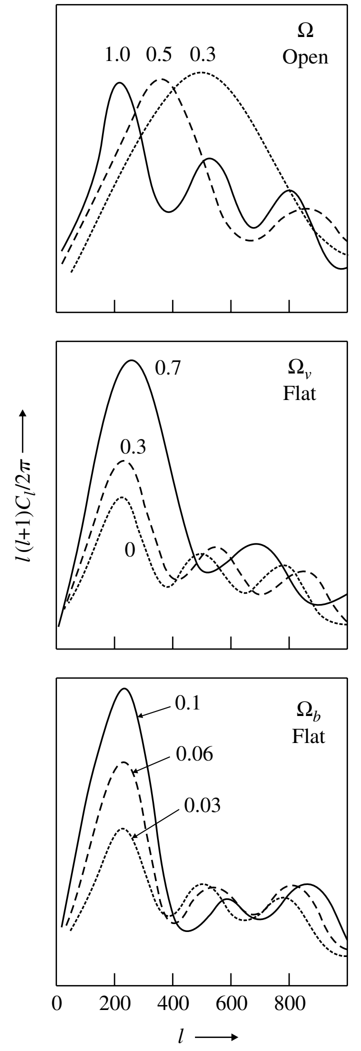


Fig. 8.9 Expected positions and heights of the 'acoustic peaks' in the angular spectrum, or l-value of the associated spherical harmonic, of the cosmic microwave radiation. In the top panel, for an open universe, the *position* (l-value) of the first peak depends on the total density Ω or the curvature $\Omega_k = 1-\Omega$. In the middle panel, for a flat universe of $\Omega = 1$, the peak position is rather insensitive to the division of density between matter and vacuum contributions, Ω_m and $\Omega_\Lambda = 1-\Omega_m$. In the bottom panel, the peak *height*—that is, the strength of the acoustic oscillation—is seen to depend on the baryon density Ω_b. An increase in baryon density raises the height of odd-numbered (compression) peaks and depresses the heights of even-numbered ones (rarefactions). The second and subsequent peaks also depend on other cosmological parameters, as described in the text. At very large l-values, the intensity of the peaks falls off exponentially because of Silk damping (after Kamionkowski and Kosowski 1999).

shorter wavelength mode which has undergone rarefaction, and so on. For much lower values of l and angles above a few degrees, the temperature variations are dominated by the SW effect described in a previous section. Because the ordinate factor $l(l+1) \sim k^2$ or $1/\lambda^2$ cancels the λ^2 dependence in (8.29), this large-angle part of the plot should be fairly flat (the so-called SW plateau). This result follows basically from scale invariance predicted in the inflation scenario and assumed in deriving (8.29).

8.14.1 Peak position versus density parameters

First we consider the case of an *open matter-dominated universe with zero cosmological constant* (i.e. $\Omega_\Lambda = 0$) at z-values where the radiation term is still comparatively small (i.e. $z < 1000$). The true coordinate distance today of an object at redshift z is $D(z)$, as given in the second of equations (5.44), so that with $\Omega_k = (1 - \Omega_m)$ and $\Omega_\Lambda = \Omega_r = 0$ we get from (5.43):

$$I(z) = \int_0^z \frac{\mathrm{d}z}{\left[\Omega_m (1+z)^3 + (1-\Omega_m)(1+z)^2\right]^{1/2}} = \int \frac{\mathrm{d}z}{\left[(1+z)(1+\Omega_m z)^{1/2}\right]} \tag{8.58}$$

This integration can be performed by making the substitution $(1 + \Omega_m z) = (1 - \Omega_m)\sec^2\theta$, when it has the value

$$\begin{aligned} I(z) &= \left[(1-\Omega_m)^{-1/2}\right] \ln\left[\frac{(1+\cos\theta)}{(1-\cos\theta)}\right] \\ &= \left(\frac{1}{q}\right) \ln\left[\frac{\{(p+q)(1-q)\}}{\{p-q)(1+q)\}}\right] \end{aligned}$$

where $p^2 = (1 + \Omega_m z)$, $q^2 = (1 - \Omega_m) = \Omega_k$. With $\sinh X = (e^X - e^{-X})/2$ the middle equation (5.44b) then gives the following expression for the distance today of an object at redshift z:

$$\begin{aligned} D(z) = R(0)r &= \left[\frac{c}{H_0 (1-\Omega_m)^{1/2}}\right] \sinh\left\{\ln\left[\frac{(p+q)(1-q)}{(p-q)(1+q)}\right]\right\} \\ &= \left(\frac{2c}{H_0}\right) \frac{\left[\Omega_m z - (2-\Omega_m)\left\{(1+\Omega_m z)^{1/2} - 1\right\}\right]}{\left[\Omega_m^2 (1+z)\right]} \end{aligned} \tag{8.59}$$

known as the *Mattig formula,* valid, we repeat, for an open, matter-dominated universe with zero vacuum energy and the radiation contribution neglected.

For a very distant object at high redshift z, (8.59) gives the simple approximate result, actually independent of the z-value, provided it is large:

$$D(z) \approx \frac{2c}{(H_0 \Omega_m)} \tag{8.60}$$

We note that for $\Omega_m = 1$, this is just the expression for the optical horizon distance in a flat, matter-dominated universe. So the effect of an open matter-dominated universe is just to enlarge the horizon by a factor $1/\Omega_m$.

The second important case is that of a *flat universe*, with the dominant contributions from *matter and from vacuum energy*, that is, with $\Omega_\Lambda = 1 - \Omega_m$ and $\Omega_k = \Omega_r = 0$. In this case the integral (5.43) has the form

$$I(z) = \int \frac{\mathrm{d}z}{\left[\Omega_m (1+z)^3 + (1-\Omega_m)\right]^{1/2}} \tag{8.61}$$

No analytical solution is possible and the integration has to be performed numerically. However, for values of $z > 4$ and $\Omega_m > 0.05$ one can neglect the vacuum term in comparison with the matter term and perform that part of the integral analytically, so that only a few minutes' work with a pocket calculator is needed to plot D versus Ω for a fixed z. We discuss the results below.

We now consider the value of the angle subtended today at the Earth by the acoustic horizon at the time of last scattering of the CMB ($t = t_{\mathrm{dec}}$). As shown in Section 5.13, matter and photon energy densities would actually have been equal at the larger value, $z_{\mathrm{eq}} \sim 5000$. But first let us assume radiation dominance for $z > z_{\mathrm{dec}}$, the epoch of last scattering. Since the curvature term is negligible in comparison with the radiation and matter terms at large z-values, distances are given by the flat universe formula in (5.44b). The ratio of sound to light velocity in a situation where radiation is prominent is $1/\sqrt{3}$ (see Table 5.2). Hence, integrating from $z = z_{\mathrm{dec}}$ to $z = \infty$, the acoustic horizon distance at the time of last scattering of the radiation is now

$$D_{\mathrm{H}} \approx \left(\frac{c}{H_0}\right) \int \frac{\mathrm{d}z}{\left[(3\Omega_r)^{1/2}(1+z)^2\right]} = \frac{c}{\left[H_0 (1+z_{\mathrm{dec}})(3\Omega_r)^{1/2}\right]} \tag{8.62}$$

In our very crude approximation, $\Omega_r(0) \sim \Omega_m(0)/(1+z_{\mathrm{dec}})$, so this result can also be written

$$D_{\mathrm{H}} \approx \frac{c}{\left[H_0 \left\{ (1+z_{\mathrm{dec}})\, 3\Omega_m \right\}^{1/2}\right]} \tag{8.63}$$

Had we instead assumed matter dominance for most of the range $z = z_{\mathrm{dec}}$ to $z = \infty$, we would have obtained exactly twice this value. If we now divide by $D(z)$ in (8.60), we obtain the value of the angle θ now subtended by the acoustic horizon at $z = z_{\mathrm{dec}}$. This therefore indicates the position of the first acoustic peak, with angle given by

$$\frac{\theta}{\sqrt{\Omega_m}} \sim [3\,(1+z)]^{-1/2} \sim 1^\circ \tag{8.64}$$

in agreement with our first estimate (8.54). This result applies for the case of an *open matter-dominated* universe with no vacuum energy term. It is based on several simplifying assumptions and the absolute value of the angle is therefore only approximate, but the result gives a feeling for the magnitudes and principal factors involved. In any case, the main result stands, namely that the angle subtended varies as $\Omega_m^{1/2}$. For the *flat universe* case, the angle has to be found by numerical integration. The results are discussed below.

As indicated by (8.64), the *position* (angle or l-value) of the first peak depends on the value of the curvature parameter k, or more specifically

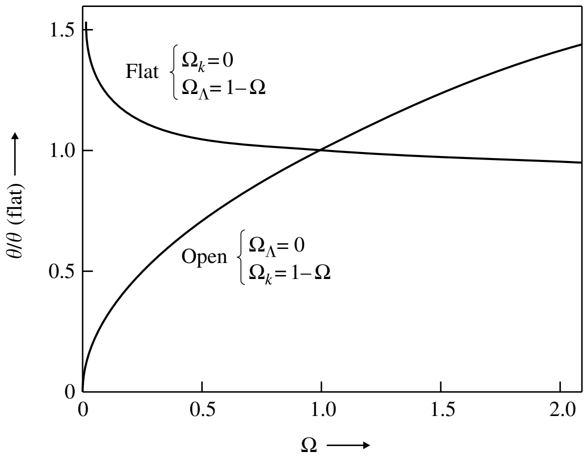


Fig. 8.10 The calculated angle subtended by the cosmic microwave anisotropy, relative to the angle expected for a flat, matter-dominated universe, plotted against the matter density parameter $\Omega = \Omega_m$. One curve is for an open universe with no vacuum energy, that is, a curvature term $\Omega_k = 1 - \Omega_m$, giving a variation proportional to $\Omega_m^{1/2}$ as in (8.64). The other curve is for a flat universe ($\Omega_k = 0$) with a vacuum energy $\Omega_\lambda = 1 - \Omega_m$. Note that the angle in this case depends rather little on the partition between matter and vacuum energy, and is close to the value for a flat matter-dominated universe with $\Omega_m = \Omega_{\text{tot}} = 1$.

$\Omega_k \equiv 1 - \Omega_m$. This is shown in Fig. 8.10 for the two important cases, of an open universe with no vacuum energy, and a flat universe with variable vacuum energy. This graph shows that, provided the total value of density parameter $\Omega_{\text{tot}} = 1$, so that the universe is flat ($\Omega_k = 0$) the position of the first peak varies little, irrespective of how the matter and vacuum energy density is shared (we remind ourselves from (5.30) that $\Omega_{\text{tot}} = \Omega_m + \Omega_\Lambda + \Omega_k + \Omega_r$ and that these symbols refer to the quantities today). Furthermore, over most of the range, the peak position is closely equal to that for a flat, matter-dominated universe, with $\Omega_m = 1$, $\Omega_\Lambda = \Omega_r = \Omega_k = 0$. It is seen that the microwave data on their own do not resolve very well the *relative* contributions from matter and vacuum energy density, and one has to appeal to other experiments to distinguish them. However, the main and very important result from the experimental results described below is that *the universe turns out to be remarkably flat.*

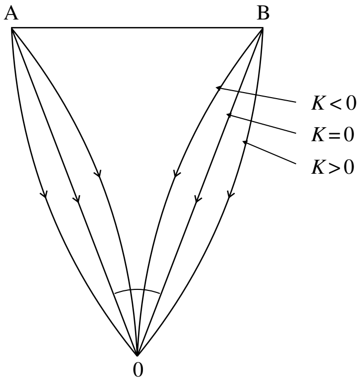


Fig. 8.11 The effect of the curvature of space (the gravitational deflection of photons) on angular measurements of distant objects. For curvature parameter $k = +1$, that is, a closed universe, the angle is increased, while for $k = -1$ (open universe) it is decreased, in comparison with the value for a spatially flat universe ($k = 0$).

In Fig. 8.11 is sketched a two-dimensional analogue of the paths of light in curved space. For a flat universe, the value of the angle θ of the peak will be approximately as in (8.64) for $\Omega_m = 1$, with the corresponding value of l in (8.56). However, for a 'closed' universe of positive curvature ($k > 0$), the angle will be increased, the gravitational field between the source and the detector acting as a converging lens. Thus the peak will move to lower l-values; while if the curvature is negative—the case of an 'open' universe—the angle will be decreased as for a diverging lens and the peak will move to higher l (see also the plots in Fig. 8.9).

As indicated in Fig. 8.9, the *height* of the first peak, and of every odd-numbered peak, measures the intensity of the acoustic compression, which is sensitive to the baryon/photon ratio (a greater baryon density will help gravitational collapse and enhance the oscillation amplitude, whereas the pressure of the photons will oppose collapse and tend to iron out inhomogeneities). The positions and heights of the other peaks are sensitive to other cosmological parameters. For example, an increase in the assumed number of neutrino flavours will push all the peaks towards higher l-values. It is of interest to remark here that the fact that there *are* in practice several peaks, kills off one model of the early universe, that of 'cosmic strings', which we therefore do not need to discuss.

8.15 Experimental observation and interpretation of CMB anisotropies

The design of experiments to detect the tiny anisotropies of order 10^{-5} in the CMB has been a very challenging problem. Because water vapour absorbs microwave radiation, the detectors have been mounted on satellites (the COBE and WMAP experiments), or flown on balloons at high and dry locations, such as the South Pole. The instruments themselves employ cryogenic cooling to reduce background, and bolometer or superconducting detectors, and may also be operated as interferometers measuring directly the tiny spatial variations in temperature. Ubiquitous background, such as infrared emission from the Milky Way, has been eliminated by means of suitable combinations of frequencies. The success of these experiments in digging out the tiny signals from the background has to be regarded as one of the great triumphs of modern experimental physics.

As mentioned above, the first crucial detection of anisotropies (the existence of which had been predicted some 30 years previously) at the 10^{-5} level was achieved by the differential microwave radiometer on the COBE satellite in 1992 (Smoot *et al.* 1992, Bennett *et al.* 1996, Mather *et al.* 2000), with, however, an angular resolution of only about 7°. Since then, measurements have been carried out with detectors having much higher angular resolution—typically $10'$–$50'$ of arc. Some were flown on high altitude balloons—the BOOMERANG (de Bernardis *et al.* 2002) and MAXIMA (Lee *et al.* 2001) experiments—and one was a ground-based interferometer—the DASI (Halverson *et al.* 2001) experiment—all at the South Pole. The most detailed data to date have been obtained with WMAP, the Wilkinson Microwave Anisotropy Probe satellite experiment (Bennett *et al.* 2003, Spergel *et al.* 2003). Over a six month period of observation, this was able to cover the entire sky. Observations were made with cooled differential radiometers at five frequency bands, ranging from 23 to 94 GHz. These variations in temperature on small scales have provided our best information on the parameters of the early universe, such as the curvature and the contributions of matter, radiation, and vacuum terms to the overall energy density.

Figure 8.12 shows the angular spectrum with its acoustic peaks in the WMAP and other experiments. The best fit to this data corresponds to the cosmic parameters shown in Table 8.1 below. There is clear evidence for a third peak, but at higher and higher l-values, the amplitude falls off exponentially because of the increasing effects of photon damping of the density fluctuations at short distances, as described in Section 8.13 above.

The microwave photons have primary anisotropies which correspond to their distribution upon leaving the 'last scattering surface' at the time of decoupling at $z_{\text{dec}} = 1100$. In fact, this is a last scattering 'shell' rather than a surface, with a thickness which is obviously of the order of the photon mean free path between Thomson scatters. If the medium were fully ionized, this is readily calculated to be $\sim$2 kpc at that time, but would be considerably more than this, since the degree of ionization was varying rapidly at $z = z_{\text{dec}}$ (recall the Saha equation in Chapter 5). So the photons last scatter over a small range of z-values (typically $\Delta z \sim 60$) and this and the effects of secondary anisotropies mean that any angular variations below about 0.03° would be completely smeared out and undetectable (see Problem 8.9).

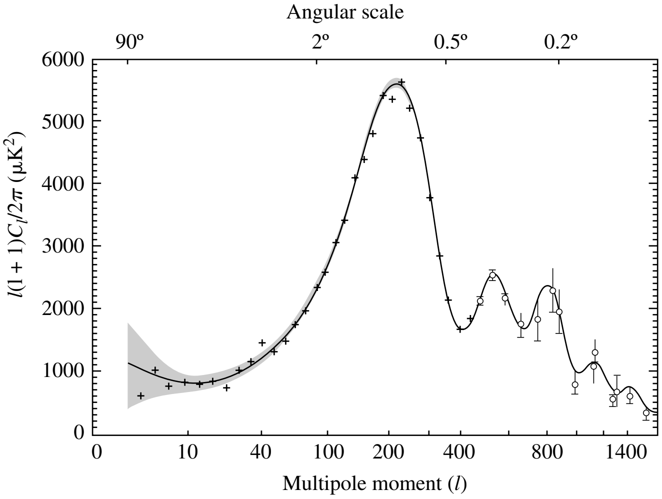


Fig. 8.12 Observed amplitude of the 'acoustic peaks' in the CMB as a function of the order l of the polynomial (8.55), from the WMAP and other experiments (Bennett *et al.* 2003). The best fit to the data—what is called the 'Standard Cold Dark Matter Model'—yields the parameters shown in Table 8.1.

Table 8.1 Best-fit cosmic parameters from combinations of measurements of CMB anisotropies, large-scale galaxy surveys and high redshift Type 1a supernova observations (from Particle Data Group, Yao *et al.* 2006)

Total closure parameter	$\Omega_{\text{tot}} = 1.00 \pm 0.02$
Dark energy contribution	$\Omega_\Lambda = 0.76 \pm 0.05$
Total matter contribution	$\Omega_m = 0.24 \pm 0.03$
Baryon density contribution	$\Omega_b = 0.042 \pm 0.004$
Hubble parameter	$H_0 = 72 \pm 3$ km s^{-1} Mpc^{-1}

The total closure density is consistent with the value close to unity predicted by the inflationary scenario. The value of the vacuum energy density or cosmological constant contribution deduced from the supernova measurements of the Hubble constant and its variation with time, described in Chapter 7, are in good agreement with estimates from combining the microwave data with galaxy redshift surveys, and of course that result, as shown in Example 5.3, gives an estimate for the age of the universe in excellent agreement with independent estimates from radioactive isotope analysis, stellar ages in globular clusters, etc. Furthermore, the baryon density in Table 8.1 is in good agreement with the (less precisely) known value from Big Bang nucleosynthesis described in Chapter 6. The bulk of the matter contribution is obviously accounted for by dark matter.

8.16 Polarization of the cosmic microwave radiation

Before decoupling from matter at $t = t_{\text{dec}}$, the microwave photons will undergo frequent Thomson scattering from the free electrons of the electron–baryon plasma. Because of the anisotropies described above, each act of scattering should result in polarization of the photons—just as light from a localized source such as the Sun becomes plane-polarized when scattered in the atmosphere, the degree of polarization perpendicular to the scattering plane being of order $\sin^2\theta$ where θ is the angle of scatter. However, since successive scatters are randomly orientated, all polarization effects will be washed out, except for that arising in the very last act of scattering before the decoupling time, t_{dec}. The direction of this polarization will vary with the angle of observation, depending on the spatial anisotropies involved. Such polarization has been measured with the DASI interferometer and more recently by the WMAP experiment, and is exactly of the magnitude (a few μK) expected. In addition to these effects at small angles of observation (but of course primarily due to large angles of scatter) polarization due to secondary processes discussed in Section 8.13.2 above, will be observed at large angles, in the region of the SW plateau, and as already described, has given information on the re-ionization of the interstellar medium.

The polarization produced in Thomson scattering is of the so-called E-mode. In general, polarization has to be described by a second-rank tensor, and can also exist in the so-called B-mode. The distinction between the two is that, if one decomposes the angular dependence into spherical harmonics, E-mode polarization has parity $(-1)^l$ while B-mode has parity (-1)$^{l+1}$. A vector interaction such as Thomson scattering cannot produce the B-mode, which requires a tensor interaction, such as that of gravity waves.

The amount of B-mode polarization due to the burst of gravitational radiation which is expected to accompany inflation, is calculated to be associated with temperature fluctuations at best at the level of 0.1 μK and will therefore be very difficult to detect. It is one of the primary aims of the future Planck satellite experiment to be launched by the European Space Agency. Of course, with the CMB photons we cannot look *directly* back before the time of matter–radiation decoupling. However, gravitational waves can be directly associated with the inflation process, and can leave their imprint on the microwave background in the form of polarization anisotropies. This seems to offer the only real hope of examining the details of the inflation process at very much earlier times, and is certainly one of the most important future goals of cosmological research.

8.17 Summary

- The conventional Big Bang model, while successfully accounting for the redshift, abundance of light elements, and the microwave background radiation, suffers from the horizon and flatness problems.
- The horizon problem arises on account of the observed isotropy of the microwave background out to the largest angles. The horizon at the time of the decoupling of matter and radiation, at $z \sim 1000$ (when the radiation had the last opportunity of interaction and achieving thermal equilibrium), now only subtends about 1° at the Earth. It is therefore impossible to understand how the large-angle uniformity in temperature could have been achieved by causal processes.
- The flatness problem arises because the fractional difference between the observed density ρ and the critical density ρ_c should be proportional to t in a radiation-dominated universe, and $t^{4/3}$ in the matter-dominated case. Thus at very early times, ρ must have been very finely tuned to ρ_c (to an accuracy of only one part in 10^{52}, if we go back to the Planck era).
- The postulate of a preliminary inflationary stage of exponential expansion, when the radius of the initial micro-universe expanded from 10^{-26} m to 1 m, solves both these problems, and also accounts for the absence of magnetic monopoles.

- Quantum fluctuations at the commencement of inflation may account for the observed perturbations, of order 10^{-5}, in the temperature (and hence density) of the cosmic microwave radiation. These quantum fluctuations would become classical, frozen fluctuations in density when they were inflated beyond the causal horizon.
- Large-scale structures in the universe—galaxies, galaxy clusters, voids, etc.—were seeded by these primordial perturbations in the density (or metric curvature).
- Initially, the density perturbations $\delta = \Delta\rho/\rho$ grew linearly with the expansion parameter R as they collapsed under gravity in the era of matter domination. The collapse became possible on all scales larger than the Jeans length, in turn determined by the speed of sound v_s in the cosmic fluid. As v_s decreased abruptly when atoms began to form, more dramatic, and non-linear, collapse on smaller and smaller scales became possible.
- The spectrum of density fluctuations predicted by inflation is a power law, with no preferred scale, and $(\Delta\rho/\rho)_{\text{rms}} \sim \lambda^{-2}$, where λ is the length involved. The spectrum observed in COBE microwave measurements at large angular scales fits this prediction. At smaller scales, below 400 Mpc, the fluctuation spectrum from galaxy surveys is much flatter, due to damping of the density perturbations by photon diffusion and collisionless damping by the relic neutrino component, and is consistent with expectations from mixed dark matter models. The degree of damping has been used to set limits of less than 0.4 eV/c^2 on the neutrino mass, summed over the three neutrino flavours.
- The tiny (10^{-5}) variations in the observed microwave temperature between pairs of points in the sky separated by angle θ can be described by a sum of Legendre polynomials $P_l(\cos\theta)$, where l-values of 100–1000 are relevant to investigations at separation angles of the order 1° or less. The amplitudes of these fluctuations in temperature, consequent on the density fluctuations, appear as a series of so-called 'acoustic peaks' when plotted against the l-value.
- The position of the first acoustic peak, at $l \approx 200$, provides a measure of the total density parameter Ω_{tot}, indicating a value close to unity, and a consequent curvature parameter $\Omega_k = 1 - \Omega$ of less than 0.05: thus the early universe is practically flat, as predicted by the inflation model. The height of the first peak and the heights and positions of the other peaks provide estimates for other cosmological parameters, such as the baryon density, the amount of dark matter, Hubble parameter, etc., and when combined with data from galaxy surveys and high redshift supernova, provide the basic parameters describing the universe.
- The CMB is found to be polarized, due to Thomson scattering from free electrons, either in a final act of scattering before decoupling at $z \sim 1000$, or later, en route through an intergalactic medium re-ionized by ultraviolet light emitted by the early stars. One also expects a separate polarization due to tensor interactions of gravitational waves accompanying inflation, and this will constitute an important test of the inflation hypothesis, hopefully to be observed by the future Planck satellite mission.

Problems

More challenging problems are denoted with an asterisk.

(8.1) Show that the free-fall time (8.34) is of the same order of magnitude as the period of a satellite in close orbit about a spherical cloud of density ρ.

(8.2) Calculate the Jeans mass and Jeans length for a mass of air at NTP (T = 273 K, $\rho = 1.29$ kg m^{-3}).

(8.3) Verify the expressions (8.42)–(8.44) for the growth of the density contrast with time in a closed ($k = +1$) matter-dominated universe. Derive the corresponding expressions for an open universe, with $k = -1$, and show that in this case the density contrast will decrease with time.

(8.4) Assume that a density fluctuation occurs during the inflation process, and that at the end of inflation, at $t_i \sim 10^{-32}$ s, this has become ‘frozen-out’ at a length scale λ, when the universe has reached a radius ~ 1 m. Calculate at what subsequent time the perturbation will come inside the horizon and start to oscillate as an ‘acoustic’ wave, for λ = 1 mm and for λ = 1 cm. Estimate the masses inside the horizon in the two cases and identify them with large-scale structures.

(8.5) Calculate the Jeans length at time t in the early, radiation-dominated universe, and show that it is approximately equal to the horizon distance at that time.

(8.6) Starting from the formula (5.44), calculate the present distance to the optical horizon D_{H} for the case of a flat universe ($k = 0$) with $\Omega_m = 0.24$ and $\Omega_\Lambda = 0.76$, neglecting the contribution of radiation. (A short numerical integration will be needed for $z < 4$. For larger z values the vacuum contribution can be neglected, and the integral evaluated analytically.)

*(8.7) Find an expression for the probability that microwave photons observed today will have undergone Thomson scattering due to re-ionization of the intergalactic medium (by ultraviolet light from the first stars), as a function of the value of the redshift below which re-ionization is assumed to take place. If this probability is 10%, find below which value of z the medium is almost totally ionized, and at what time after the Big Bang this occurs. Assume that the universe has matter energy density with $\Omega_m = 0.26$ and baryon density $\Omega_b = 0.045$, and for simplicity suppose that for the z-values involved, radiation, vacuum energy, and curvature can all be neglected (i.e. $\Omega_r = \Omega_\Lambda = \Omega_k = 0$).

(8.8) Estimate the escape velocities from a typical galaxy and from a typical galaxy cluster. Compare this with the mean velocity of relic neutrinos, assuming that they are non-relativistic with a Maxwellian distribution in velocity, and have masses of 0.1, 1.0, or 10 eV/c^2.

*(8.9) Calculate the mean path length between Thomson scatters of the CMB photons at $z \sim z_{\mathrm{dec}} = 1100$, assuming the medium to be completely ionized. Because the appropriate degree of ionization near the point of decoupling is small, assume the thickness of the ‘last scattering shell’ is 10 times this length. Calculate the corresponding variation in Δz and the resultant angular smearing of the acoustic peaks in the angular spectrum of the radiation. (*Hint*: refer to (8.64) and to (5.43) and (5.44), assuming a flat universe).

Part 3

Particles and Radiation in the Cosmos

This page intentionally left blank

Cosmic particles

9

9.1 Preamble

The particles circulating in the cosmos include the so-called cosmic rays, which have been intensively studied ever since their discovery by Hess in 1912. Karl K. Darrow, a past chairman of the American Physical Society, caught some of the atmosphere of this early research when he described their study as remarkable ‘for the delicacy of the apparatus, the minuteness of the phenomena, the adventurous excursions of the experimenters and the grandeur of the inferences’.

Cosmic rays consist of high-energy particles incident on the Earth from outer space, plus the secondary particles which they generate as they traverse the atmosphere. Their study has a special place in physics, not only in its own right, but because of the pioneering role that cosmic ray research has played—and is still playing—in the study of elementary particles and their interactions. We can recall the discovery in cosmic rays of antimatter, in the form of the positron and e^+e^- pair production in 1932, and of pions and muons and strange particles in the late 1940s. In those days, before 1950, the cosmic radiation was the only available source of high-energy particles (those above about 1 GeV). These discoveries indeed kick-started the building of large particle accelerators and the development of their associated detecting equipment, developments which were essential in broadening the scope of the subject and putting elementary particle physics on a sound quantitative basis.

Later, in the 1980s and 1990s, the study of the interactions of solar and atmospheric neutrinos, on distance scales far larger than anything that had been attempted at accelerators or reactors, revealed the first cracks in the Standard Model, with evidence for neutrino flavour mixing and for finite neutrino masses, as described in Section 4.2 and Sections 9.15–9.17. This has led to a revival, in the new millennium, of lepton physics in fixed-target experiments at accelerators, and to the development of radically new proposals such as the building of muon storage rings to serve as sources of high-energy electron– and muon–neutrinos.

At the highest energies, study of γ-rays in the TeV range and above has indicated point sources in the skies where it seems the most violent events in the universe have taken place, and intensive studies of both γ-rays and ultra-high-energy protons and heavier nuclei will certainly shed new light on mechanisms for particle acceleration, as well perhaps as revealing new fundamental processes taking place at energies far in excess of what could ever be achieved on the Earth. The study of cosmic rays is indeed still a very open field, where new developments and mysteries arise on an almost daily basis.

9.2 The composition and spectrum of cosmic rays

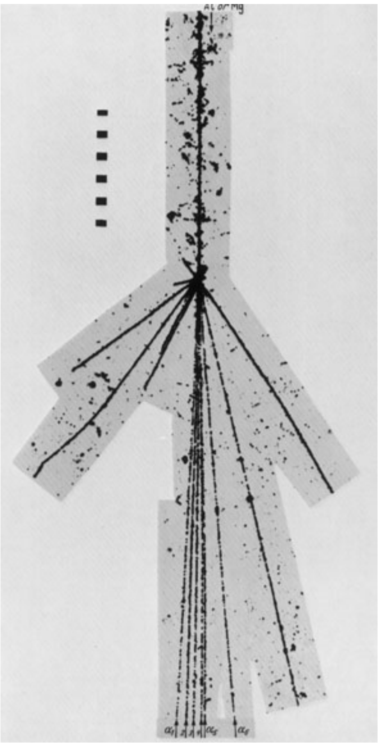


Fig. 9.1 Photomicrograph of a track due to an incident high-energy primary cosmic ray nucleus of aluminium ($Z = 13$) which interacts in a 'nuclear' photographic emulsion flown on a balloon in the stratosphere. As explained in Chapter 6, the emulsion consists of a suspension of microcrystals of silver bromide or iodide (of order 0.25 μm in radius) in gelatine. Charged particles ionize the atoms they traverse and after processing, the unaffected halide is dissolved out and the affected microcrystals are reduced to black metallic silver, so forming the tracks. In this example, the incident nucleus undergoes fragmentation into a 'jet' of six alpha particles ($Z = 2$) in the forward direction. The charges of primary and secondary nuclei were measured from the frequency along the tracks of the hair-like δ-rays (knock-on electrons), which varies as Z^2. The tracks at wider angles are due to protons ejected from the struck nucleus. The mean free path for interaction of heavy primary nuclei with those of the atmosphere is of order 10 gm cm^{-2} so that they do not penetrate much below an altitude of 25 km. The scale on the left-hand side of this photomicrograph is 50 μm.

The charged primary particles of the cosmic rays consist principally of protons (86%), alpha particles (11%), nuclei of heavier elements up to uranium (1%), and electrons (2%). While these come from primary sources, there are also very small proportions of positrons and antiprotons, which we believe are of secondary origin and generated by interactions of the primary particles with interstellar gas. The above percentages are for particles above a given *magnetic rigidity* $R = pc/z\,|e|$, where p is the momentum and $z|e|$ the particle charge, that is, for particles with the same probability of penetrating to the atmosphere through the geomagnetic field. Neutral particles consist of γ-rays and of neutrinos and antineutrinos. Some of these can be identified as coming from 'point' sources in the sky; for example, neutrinos from the Sun and from supernovae and γ-rays from sources such as the Crab Nebula and active galactic nuclei (AGNs).

The nature of the charged primary particles was first established by means of nuclear emulsion detectors flown in high altitude balloons—see Fig. 9.1. At low energies, primary energies may be estimated from the range in absorber. For the GeV–TeV energy region, the calorimetric method has been employed. This involves measuring the ionization energy in electromagnetic showers which develop as a result of the nuclear cascade which the primary generates as it traverses great thicknesses of absorber (see Section 9.7). Detectors flown in satellites have employed scintillation counters to measure the primary nuclear charge from the pulse height, and gas-filled Cerenkov counters to measure the particle velocity and hence the energy (see Fig. 9.9).

The energy density in the cosmic rays is about 1 eV cm^{-3} when calculated for deep space outside the influence of solar system magnetic fields, and is therefore quite comparable with the energy density in starlight of 0.6 eV cm^{-3}, that in the cosmic microwave background radiation of 0.26 eV cm^{-3}, and that in the galactic magnetic field of 3 μG or 0.25 eV cm^{-3}. The galactic magnetic fields are mostly trapped inside the spiral arms of the galaxy. Galactic and intergalactic magnetic fields are discussed in Section 8.2.

The bulk of the primary radiation is of galactic origin: however, the fact that the spectrum extends to very high energies (over 10^{20} eV) indicates that some at least of the radiation could be of extra-galactic origin, since the galactic magnetic field could not contain such particles inside our local galaxy. Indeed, as indicated below, the spectrum hardens at the so-called ankle at about 4×10^{18} eV, perhaps suggestive of an extra-galactic source.

Example 9.1 *Calculate the radius of curvature of the trajectory of a proton of energy 10^{20} eV in an (assumedly uniform) galactic magnetic field of 3 microgauss (3×10^{-10} Tesla). Compare this with the typical disc thickness of a spiral galaxy of $d = 0.3$ kpc.*

The radius of curvature ρ in metres of a singly charged particle of momentum p GeV/c in a magnetic field of B tesla is given by $\rho = pc/(0.3B)$—see Appendix A. Substituting one finds $\rho = 10^{21}$ m or 36 kpc. So the magnetic deflection is only $d/\rho \sim 0.5°$. If the extra-galactic field is small, it is therefore feasible to search for possible point sources of such high-energy particles (see Section 9.13).

The chemical composition of the cosmic ray nuclei exhibits remarkable similarities to the solar system abundances, which are deduced from absorption lines in the solar photosphere and from meteorites, but it also shows some significant differences, as seen in Fig. 9.2. The cosmic and solar abundances both show the odd–even effect, associated with the fact that nuclei with Z and A even are more strongly bound than those with odd A and/or odd Z, and therefore are more frequent products in thermonuclear reactions in stars. The peaks in the normalized abundances for C, N, and O, and for Fe are also closely similar, suggesting that many of the cosmic ray nuclei must be of stellar origin.

The big differences between the cosmic and solar abundances are in those of Li, Be, and B. The abundance of such elements in stars is very small, since they have low Coulomb barriers and are weakly bound and rapidly consumed in

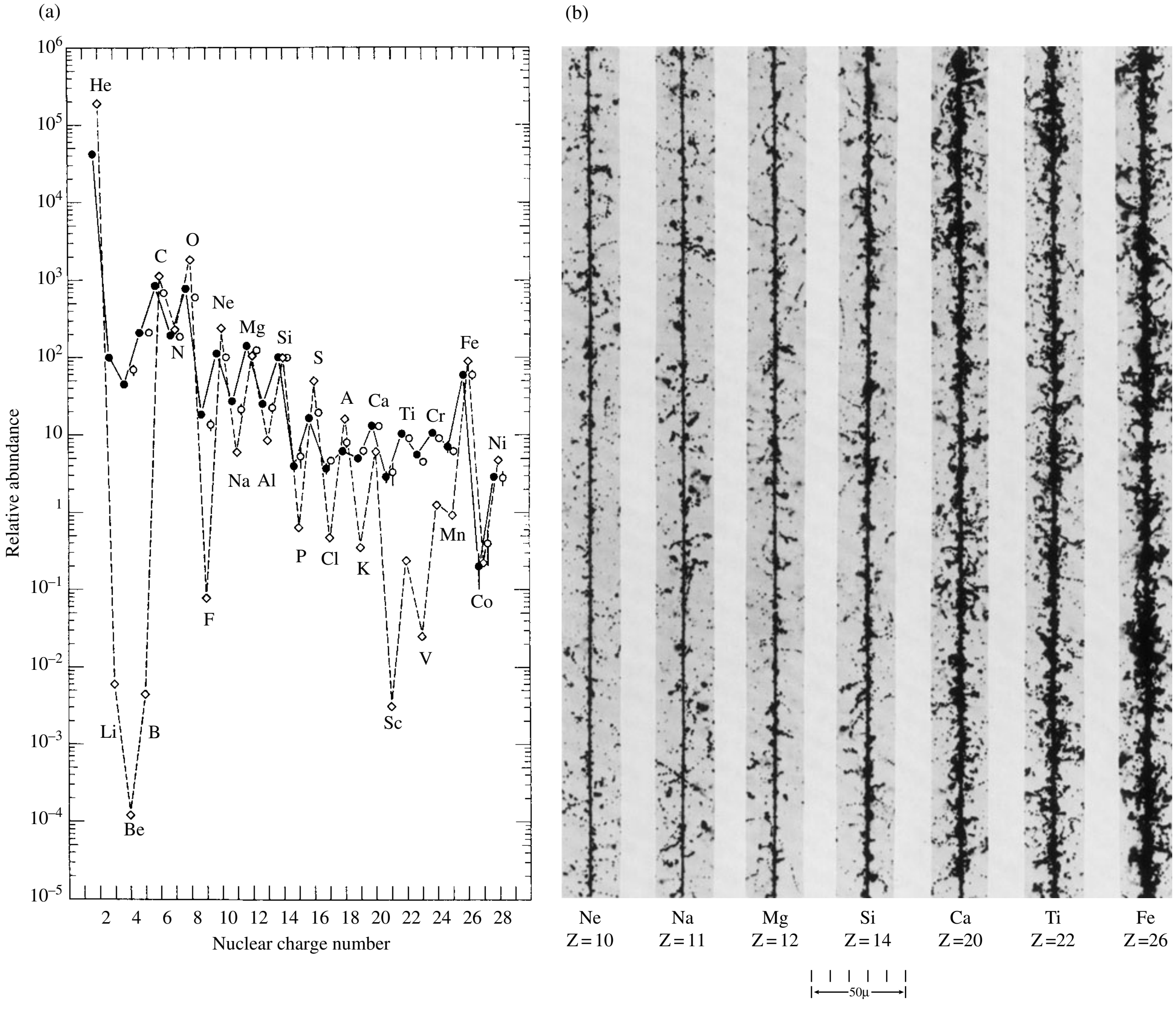


Fig. 9.2 (a) The chemical composition of primary cosmic ray nuclei, shown full-line, compared with the solar abundances of the elements, shown as a dashed line (from Simpson (1983) with permission from *Annual Reviews of Nuclear and Particle Science* Vol. 33). (b) Tracks of various primary cosmic ray nuclei recorded in nuclear emulsions flown on a high altitude balloon.

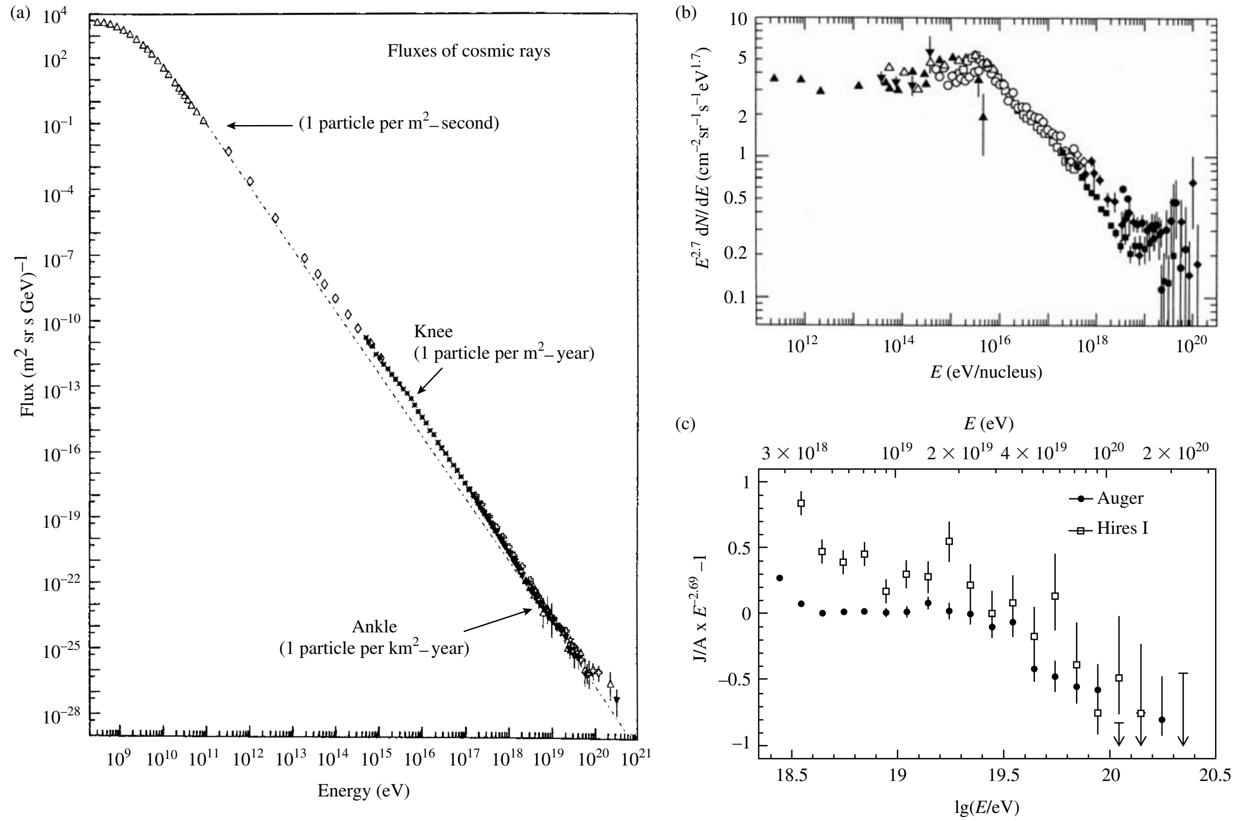


Fig. 9.3 (a) The primary cosmic ray spectrum showing the power law $E^{-2.7}$ dependence at energies below the 'knee', steepening to $E^{-3.0}$ at energies above it, followed by indication of a flattening above the 'ankle' at $\sim 4 \times 10^{18}$ eV. Arrows show the integral fluxes of particles above certain energies (graph by S. Swordy, reproduced courtesy of J.W. Cronin (1999)). (b) The primary spectrum multiplied by $E^{2.7}$, showing the knee in more detail (from *Review of Particle Properties*, by Barnett *et al.* 1996). (c) The detail of the spectrum at the very highest energies, from the AUGER and HiRes extensive air shower arrays (see Section 9.12). The vertical scale shows the fractional difference between the observed spectrum and one with a flux varying as $E^{-2.69}$. The spectrum hardens beyond the 'ankle' at 4×10^{18} eV, followed by the (presumed) GZK cut-off above 4×10^{19} eV. Error bars indicate statistical errors (reproduced courtesy of AUGER collaboration).

nuclear reactions in stellar cores. Their comparative abundance in cosmic rays is due to *spallation* of carbon and oxygen nuclei as they traverse the interstellar hydrogen (see Fig. 9.1). In fact the amount of these light elements determines the average thickness of interstellar matter which the radiation traverses and indicates an average lifetime of the cosmic rays in the galaxy of about 3 million years. It is found that the energy spectra of Li, Be, and B are somewhat steeper than those of carbon or oxygen, indicating that at the higher energies nuclei do not undergo so much fragmentation, presumably because they leak out of the galaxy sooner than those of lower energy. In a similar way, the abundance of Sc, Ti, V, and Mn in the cosmic rays is due to spallation of the abundant Fe and Ni nuclei.

Figure 9.3 shows the energy spectrum of cosmic ray protons. Above a few GeV energy, the spectrum up to the so-called knee at 10^{16} eV (10^4 TeV) follows a simple power law

$$N(E)\,\mathrm{d}E = \mathrm{const} \cdot E^{-2.7}\,\mathrm{d}E \qquad E < E_{\mathrm{knee}} = 10^{16}\,\mathrm{eV} \tag{9.1}$$

Above this 'knee' the spectrum becomes steeper with an index of approximately −3.0

$$N\,(E)\,\mathrm{d}E = \mathrm{const}\cdot E^{-3.0}\,\mathrm{d}E \quad E_{\mathrm{ankle}} > E > E_{\mathrm{knee}} \tag{9.2a}$$

before hardening again above the so-called ankle at $E_{\mathrm{ankle}} \approx 4 \times 10^{18}$ eV:

$$N\,(E)\,\mathrm{d}E = \mathrm{const}\cdot E^{-2.69}\,\mathrm{d}E \quad E_{\mathrm{GZK}} > E > E_{\mathrm{ankle}} \tag{9.2b}$$

Above $E_{\mathrm{GZK}} = 4 \times 10^{19}$ eV, the spectra found by extensive air shower experiments—the AUGER experiment in Argentina and the HiRes ('Fly's Eye') detector in Utah—appear to fall off, presumably because of the 'GZK cut-off' due to pion production in collisions with the microwave background photons (see Section 9.12). They are parameterized by the form

$$N\,(E)\,\mathrm{d}E = \mathrm{const}\cdot E^{-4.2}\,\mathrm{d}E \quad E > E_{\mathrm{GZK}} = 4 \times 10^{19}\,\mathrm{eV} \tag{9.2c}$$

From the development of the air showers as a function of atmospheric depth, it is known that over the entire energy ranges above, the primary particles are both protons and heavier nuclei.

At energies above 30 GeV, where effects due to the magnetic fields of the Earth or the Sun are unimportant, the radiation appears to be isotropic, since the galactic magnetic fields would destroy any initial anisotropy except at extremely high energies. Data from the AUGER air shower experiment does detect anisotropies and close and significant correlations of showers above 6×10^{19} eV with known AGNs within about 75 Mpc of the Earth (see Section 9.13).

9.3 Geomagnetic and solar effects

The primary radiation, for charged particles below 10 GeV energy, does show directional effects and also time dependence. The charged primaries are affected by the Earth's magnetic field, which approximates that due to a simple magnetic dipole, and also by modulation in time due to the solar wind, which follows the 11-year solar cycle.

We first discuss the geomagnetic effects. The axis of the dipole is at an angle to the axis of the Earth's rotation. The geographical coordinates of the poles varies slowly with geological time, the present N pole being located at longitude 101° W, latitude 75° N. The calculation of the actual orbits of particles incident on the Earth as they spiral in the dipole field is rather tedious and complex, and most easily accomplished using a computer program. However, some of the main features of the geomagnetic effects can be understood analytically.

Consider first a particle of charge $z|e|$, velocity v, and momentum $p = mv$ travelling in a circular equatorial path of radius r around a short dipole of moment M. Equating centrifugal and magnetic forces we obtain

$$z\,|e|\,|\mathbf{B} \times \mathbf{v}| = \frac{mv^2}{r}$$

where the equatorial field due to the dipole is

$$B = \left(\frac{\mu_0}{4\pi}\right)\frac{M}{r^3}$$

The radius of the orbit is therefore

$$r_S = \left[\left(\frac{\mu_0}{4\pi}\right)\frac{Mz\,|e|}{p}\right]^{1/2} \tag{9.3}$$

known as the Stɸrmer unit, after the physicist who first treated the problem. A significant value of the particle momentum is that which makes the Earth's radius r_E equal to one Stɸrmer unit, that is,

$$\frac{pc}{z} = \left(\frac{\mu_0}{4\pi}\right)\frac{Mc\,|e|}{r_E^2} = 59.6 \text{ GeV} \tag{9.4}$$

where we have inserted the values in SI units of $\mu_0/4\pi = 10^{-7}, M = 8 \times 10^{22}$ amp m, $r_E = 6.38 \times 10^6$ m, $|e| = 1.6 \times 10^{-19}$ coulomb, and 1 GeV $= 1.6 \times 10^{-10}$ J. A simple construction makes clear that no proton of momentum less than the above value can reach the Earth from the eastern horizon at the magnetic equator. Stɸrmer showed that the equation of motion obeyed by a particle has the form

$$b = r \sin\theta \cos\lambda + \frac{\cos^2\lambda}{r} \tag{9.5}$$

where r is the distance of the particle from the dipole centre in Stɸrmer units, λ is the geomagnetic latitude, and θ is the angle between the velocity vector **v** and its projection in the meridian plane OAB co-moving with the particle—see Fig. 9.4. The angle θ is called positive for particles travelling from east to west, as that shown, while it is negative for particles travelling in the opposite direction. The quantity b (again in Stɸrmer units) is the impact parameter or closest distance of approach to the dipole axis by a tangent to the particle trajectory at infinity. Since we must have $|\sin\theta| < 1$, (9.5) places restrictions

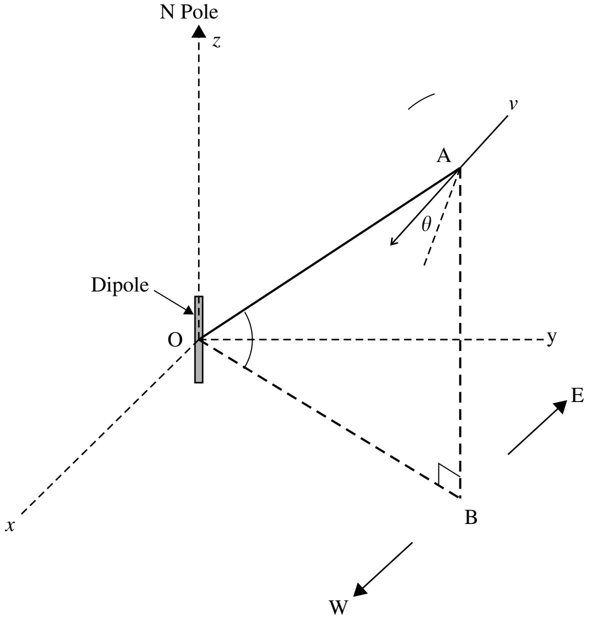


Fig. 9.4 Coordinate system and variables describing a particle A with velocity **v** in the field of a dipole M at O. θ is the angle between the velocity vector **v** of the particle and the meridian plane OAB rotating with the particle.

on the values of b, r, and λ for the 'allowed' trajectories of particles reaching the Earth. The condition $b \leq 2$ is found to be critical in determining which momenta are cut off by the Earth's field. Inserting $b = 2$ in (9.5) the equation for the cut-off momentum at any λ and θ is given by

$$r = \frac{\cos^2 \lambda}{\left[1 + \left(1 - \sin\theta \cos^3 \lambda\right)^{1/2}\right]} \tag{9.6a}$$

where from (9.4)

$$\frac{pc}{z} = 59.6\, r^2 \text{ GeV} \tag{9.6b}$$

since we are concerned with particles arriving at the Earth, so that $r = r_E/r_S$. For example, for particles incident from the vertical, $\theta = 0$ and $r = \frac{1}{2}\cos^2 \lambda$ so that the cut-off momentum is

$$(pc)_{\text{min}}\,(\theta = 0) = 14.9z \cos^4 \lambda \text{ GeV} \tag{9.7}$$

In NW Europe, for example, with $\lambda \sim 50°$ N, the vertical cut-off momentum would be $pc/z = 1.1$ GeV, or a minimum kinetic energy for a proton of 0.48 GeV.

From (9.6) and (9.7) we see that at the magnetic equator, the vertical cut-off is 14.9 GeV/c. That for particles from the eastern horizon ($\sin\theta = +1$) is 59.6 GeV/c while that for particles from the western horizon ($\sin\theta = -1$) is only $59.6/\left(1 + \sqrt{2}\right)^2 = 10.2$ GeV/c. This results in the so-called *east–west effect,* namely that at all latitudes, more (positively charged) particles arrive from the west than from the east, because of the lower momentum cut-off. The effect. arises essentially because all positively charged particles are deflected in a clockwise spiral, as viewed from above the N pole. Figure 9.5 shows a map of the vertical cut-off rigidities.

The azimuthal and latitude dependencies of the primary particles are promulgated in the secondaries they produce in traversing the atmosphere. Such effects are observed, for example, in the interactions of atmospheric neutrinos, coming from the decay of secondary pions and muons, and were important in establishing the credibility of the experiments and their interpretation in terms of neutrino flavour oscillations, as discussed in Section 9.15.

Example 9.2 *Estimate the ratio of intensity of primary protons incident from the eastern horizon, as compared with that from the west, at magnetic latitude 45° N.*

From (9.6) the cut-off momentum is

$$\frac{59.6 \cos^4 \lambda}{\left[1 + \left(1 - \cos^3 \lambda\right)^{1/2}\right]^2} = 4.58\ \frac{\text{GeV}}{\text{c}}$$

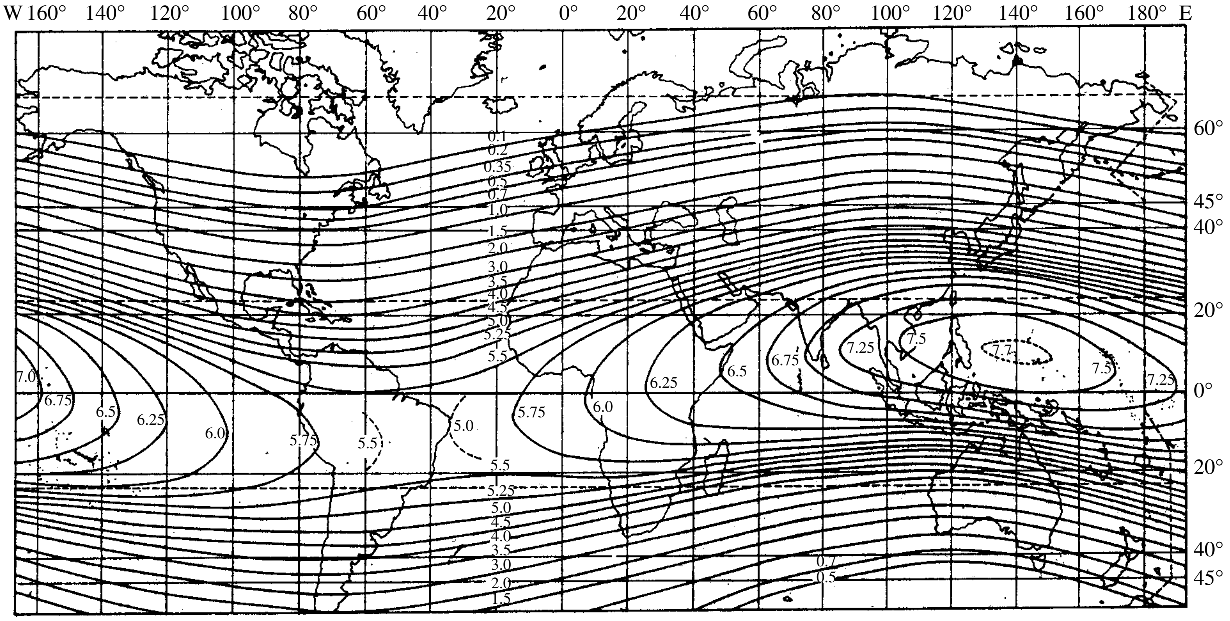


Fig. 9.5 Map of vertical geomagnetic cut-off values, given as kinetic energy in GeV per nucleon, for nuclei with $A = 2Z$. The values were calculated for a displaced dipole field. The maximum cut-off is about 7.7 GeV, or a momentum cut-off of 8.6 GeV/c, per nucleon. For protons the momentum cut-off would then be 17.2 GeV/c, to be compared with the value (9.7) for an undisplaced dipole field (from Webber 1958).

from the east and

$$\frac{59.6\cos^4\lambda}{\left[1+\left(1+\cos^3\lambda\right)^{1/2}\right]^2} = 3.18\ \frac{\text{GeV}}{\text{c}}$$

from the west. Assuming a power law momentum spectrum of the form $\mathrm{d}p/p^{2.7}$, the ratio of eastern over western intensities is $(3.18/4.58)^{1.7} = 0.54$.

In reality, the Earth's magnetic dipole (formed by ring currents deep in the Earth) is offset by some 400 km from the Earth's centre, and there are also higher-order (quadrupole) components to the field. Furthermore, at distances beyond a few Earth's radii, the trajectories are strongly distorted by the effects of the *solar wind*, which is a plasma of low-energy protons and electrons ejected from the Sun. Variations in this wind follow the 11-year sunspot cycle. The counting rate of sea-level neutron monitors has been measured for many decades and is in exact anti-correlation with the sunspot number, the difference between maximum and minimum counting rates being of order 20%. Although the protons and electrons in the solar wind are of low energy (with proton kinetic energies of order 0.5 keV) they have high intensities, with a kinetic energy density of order 3 keV cm^{-3} and an associated magnetic field of about $10^{-8}T$. If the Earth happens to be in the path of this wind, it experiences phenomena called *solar flares*. For example, dramatic aurora phenomena are observed in latitudes near the magnetic poles. These arise because charged particles from the flare become scattered and trapped into the Earth's field (which acts as a sort of magnetic mirror), spiralling to and fro from pole to pole around the lines of force and producing excitation of the air molecules in the stratosphere, with the resultant optical display.

9.4 Acceleration of cosmic rays

How do cosmic rays attain their colossal energies, up to at least 10^{20} eV, and how do we account for the form of the energy spectrum? Many years ago, it was remarked that the energy density in cosmic rays, coupled with their lifetime in the galaxy, required a power supply somewhat similar to the rate of energy generation in supernova shells. Our own galaxy has a radius $R \sim 15$ kpc and disc thickness $D \sim 0.3$ kpc. The total power requirement to accelerate the cosmic rays in the disc, for an average energy density of $\rho_E = 1\,\text{eV cm}^{-3}$ is thus

$$W_{\text{CR}} = \frac{\rho_{\text{E}}\pi R^2 D}{\tau} = 3 \times 10^{41}\ \text{Jyr}^{-1} \tag{9.8}$$

where $\tau \sim 3$ million years is the average age of a cosmic ray particle in the galaxy, before it diffuses out or is depleted and lost in interactions with the interstellar gas. A Type II supernova (see Section 10.8) typically ejects a shell of material of about 10 solar masses (2×10^{31} kg), with velocity of order 10^7 m s^{-1} into the interstellar medium, at a rate based on an average over many galaxies of around 2 ± 1 per century. (In our galaxy in fact only eight have been reported in the last 2000 years.) This gives an average power output per galaxy of

$$W_{\text{SN}} = 10^{43}\ \text{J yr}^{-1} \tag{9.9}$$

Although the galactic supernova rate is somewhat uncertain, it appears therefore that an efficiency for the shock-wave to transmit energy to cosmic rays of a few percent would be enough to account for the total energy in the cosmic ray beam.

In the 1950s, Fermi had considered the problem of cosmic ray acceleration. He first envisaged charged cosmic ray particles being reflected from ‘magnetic mirrors’ provided by the fields associated with massive clouds of ionized interstellar gas in random motion. It turns out, however, that such a mechanism is too slow to obtain high particle energies in the known lifetime of cosmic rays in the galaxy. Fermi also proposed that acceleration could occur due to *shock fronts*. Consider, in a simplified one-dimensional picture (Fig. 9.6) a relativistic particle travelling in the positive x-direction, which traverses a shock front moving with velocity $-u_1$ in the negative x-direction. Suppose that the particle is back-scattered by the field in the gas behind the front, which will have a

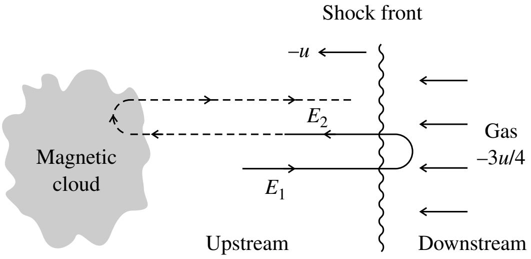


Fig. 9.6 Diagram depicting acceleration of a charged particle on crossing a shock front, and being scattered back across the front by the upstream gas.

velocity component in the direction of the shock of

$$u_2 = \frac{2u_1}{\left(C_p/C_v + 1\right)} = \frac{3u_1}{4} \tag{9.10}$$

where the ratio of specific heats $C_p/C_v = 5/3$ for an ionized gas. Thus the particle travels back with velocity u_2 across the shock front, to be scattered by magnetized clouds upstream of the front. If these again scatter the particle backwards (that is, in the direction of positive x), the particle can re-cross the front and repeat the cycle of acceleration once more. Because the front is planar (i.e. unidirectional) a straightforward application of the Lorentz transformations (see Chapter 2) shows that the fractional energy gain is of the order of the shock front velocity (see Problem 9.11):

$$\frac{\Delta E}{E} \sim \frac{u_1}{c} \tag{9.11}$$

There are many possible sources of shocks, but as indicated above, Type II supernovae shells seem to be good candidates, with shock velocities of order 10^7 m s^{-1}. Suppose now that, in each cycle of acceleration at the shock front, the particle gets an energy increment $\Delta E = \alpha E$. After n cycles its energy becomes

$$E = E_0 \left(1 + \alpha\right)^n$$

Thus in terms of the final energy the number of acceleration cycles is

$$n = \frac{\ln\left(E/E_0\right)}{\ln\left(1 + \alpha\right)} \tag{9.12}$$

At each stage of the acceleration the particle can escape further cycles. Let P be the probability that the particle stays for further acceleration, so that after n cycles the number of particles remaining for further acceleration will be

$$N = N_0 P^n$$

where N_0 is the initial number of particles. Substituting for n we get

$$\ln\frac{N}{N_0} = n \ln P = \ln\left(\frac{E}{E_0}\right)\frac{\ln P}{\ln\left(1 + \alpha\right)} = \ln\left(\frac{E_0}{E}\right)^s$$

where $s = -\ln P/\ln\left(1 + \alpha\right)$. The number N will be the number of particles with n or more cycles, thus with energy $\geq E$. Hence the differential energy spectrum will follow the power law dependence

$$\frac{\mathrm{d}N\left(E\right)}{\mathrm{d}E} = \text{constant} \times \left(\frac{E_0}{E}\right)^{(1+s)} \tag{9.13}$$

For shock-wave acceleration, it turns out that $s \sim 1.1$ typically, so that the differential spectrum index is -2.1, compared with the observed value of -2.7. The steeper observed spectrum could be accounted for if the escape probability $(1-P)$ was energy dependent. As we have already seen, the spallation spectrum of Li, Be and B indeed falls off more rapidly with energy than that of the parent

C and O nuclei, indicating that the escape probability does indeed increase with energy.

The shock-wave acceleration from supernovae shells appears capable of accounting for the energies of cosmic ray nuclei of charge $z|e|$ up to about $100z$ TeV ($10^{14}z$ eV), but hardly beyond this. Other mechanisms must be invoked for the very-highest-energy cosmic rays, and among the processes likely to play an important part are those associated with accretion of matter from nearby stars and gas on to massive black holes at the centre of AGNs. This is supported by data on correlations with AGNs for the most energetic particles, as described in Section 9.13. The enormous tidal forces involved mean that particles in the rapidly spinning accretion discs can be accelerated to tangential velocities approaching light velocity. However, the detailed mechanisms involved are not presently understood.

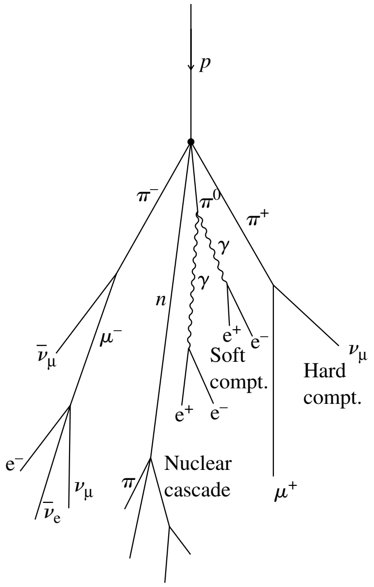


Fig. 9.7 Diagram (not to scale) indicating the production and decay of pions and muons in the atmosphere.

9.5 Secondary cosmic radiation: pions and muons—hard and soft components

The term 'cosmic rays' properly refers to particles and radiation incident from outside the Earth's atmosphere. These primary particles will produce secondaries (mesons) in traversing the atmosphere, which plays the same role as a target in an accelerator beam. The situation is shown schematically in Fig. 9.7. The most commonly produced particles are pions, which occur in three charged and neutral states π^+, π^-, and π^0. Since the nuclear interaction mean free path in air is $\lambda_{\text{int}} \sim 100$ gm cm^{-2} for a proton (and much less for a heavy nuclear primary), compared with a total atmospheric depth of $X = 1030$ gm cm^{-2}, the pions are created mostly in the stratosphere. The *charged pions* decay to muons and neutrinos: $\pi^+ \to \mu^+ + \nu_\mu$ and $\pi^- \to \mu^- + \bar{\nu}_\mu$, with a proper lifetime of $\tau = 26$ ns and a mean free path before decay of $\lambda_{\text{dec}} = \gamma c \tau$ where $\gamma = E_\pi / m_\pi c^2$ is the time dilation factor. With $m_\pi c^2 = 0.139$ GeV, $\lambda_{\text{dec}} = 55$ m for a 1 GeV pion. To a rough approximation the upper atmosphere is isothermal, and the depth x (gm cm^{-2}) then varies exponentially with height h (kms), according to the formula

$$x = X \exp\left(\frac{-h}{H}\right) \quad \text{where } H = 6.5\,\text{km}. \tag{9.14}$$

Differentiating this expression, one sees that in an interval Δh of $\lambda = 55$ m $\sim 0.01H$, the depth will change by only 1%. Thus nuclear absorption will only become important for charged pions with $\lambda \sim H$ or energies of 100 GeV or more. At GeV energies practically all charged pions decay in flight (rather than interact).

The daughter muons are also unstable, undergoing the decay $\mu^+ \to \text{e}^+ + \nu_e + \bar{\nu}_\mu$, with a proper lifetime of $\tau = 2200$ ns. Since the muon mass is 0.105 GeV, a one GeV muon has a mean decay length of 6.6 km, about equal to the scale height H of the atmosphere. Muons of energy 1 GeV or less will therefore decay in flight in the atmosphere (there is no competition with nuclear interaction since muons do not have strong interactions). However, a 3 GeV muon, for example, has a mean decay length of 20 km, of the same order as the typical distance from its point of production to sea-level. Moreover, with an ionization energy loss rate of 2 MeV gm^{-1} cm^2 of air traversed (see (9.15)), muons with 3 GeV or more energy can get through the entire atmosphere without being brought to rest or decaying. Still higher-energy muons can reach deep underground, and for this reason they are said to constitute the *hard component* of the cosmic radiation. The remaining products of charged pion and muon production, the neutrinos, are discussed in Section 9.15.

The *neutral pions* undergo electromagnetic decay, $\pi^0 \rightarrow 2\gamma$, with an extremely short lifetime of 8×10^{-17} s. The photons from the decay develop electron–photon cascades, as described below, mostly in the high atmosphere, since the absorption length of these cascades is short compared with the total atmospheric depth. The electrons and photons of these cascades constitute the easily absorbed *soft component* of the cosmic radiation.

Among the products of the nuclear interactions of primary cosmic rays in the atmosphere are radioactive isotopes, of which an important one is ^{14}C formed, for example, by neutron capture in nitrogen: $n+^{14}\mathrm{N} \rightarrow^{14} \mathrm{C}+^{1}\mathrm{H}$. The ^{14}C atoms produced in this way combine to form CO_2 molecules and thus participate, like the more common, stable ^{12}C atoms, in the circulation of this gas in the atmosphere, through rainfall into the oceans, and in absorption in organic matter. Since carbon-14 has a mean lifetime of 5600 years, its abundance relative to carbon-12 in organic matter can be used to date the sample. This of course assumes that the carbon-14 production rate by cosmic rays has been constant with time. In fact, comparison of the age from the isotope ratio with that from ancient tree ring counts shows that the cosmic ray intensity did vary in the past and was some 20% larger 5000 years ago. This variation was presumably due to long-term fluctuations in the value of the Earth's magnetic field which, associated with continental drift, is known from rock samples to have changed its sign and magnitude many times over geological time.

9.6 Passage of charged particles and radiation through matter

As a preliminary to discussing the subjects of γ-ray sources and extensive air showers, we include here a summary of the interactions of charged particles and radiation as they traverse matter, in solid materials and in the atmosphere.

9.6.1 Ionization energy losses

Charged particles lose energy as a result of collisions with atomic electrons, leading to ionization of the atoms. The rate of *ionization energy loss* is given by the Bethe–Bloch formula

$$\left(\frac{\mathrm{d}E}{\mathrm{d}x}\right)_{\mathrm{ion}} = \left(\frac{4\pi N_0 z^2 e^4}{mv^2}\right)\left(\frac{Z}{A}\right)\left\{\ln\left[\frac{2mv^2\gamma^2}{I}\right] - \beta^2\right\} \tag{9.15}$$

where m is the electron mass and v and ze are the velocity and charge of the incident particle, $\beta = v/c, \gamma^2 = 1/\left(1-\beta^2\right)$, N_0 is Avogadro's number, Z and A are the atomic and mass numbers of the atoms of the medium, and x is the path length in the medium, usually measured in g cm^{-2}. The quantity I is the mean ionization potential of the medium, averaged over all electrons in the atom, and is approximately $I \approx 10\,\mathrm{Z}\,\mathrm{eV}$. Notice that $\mathrm{d}E/\mathrm{d}x$ is a function of velocity v and

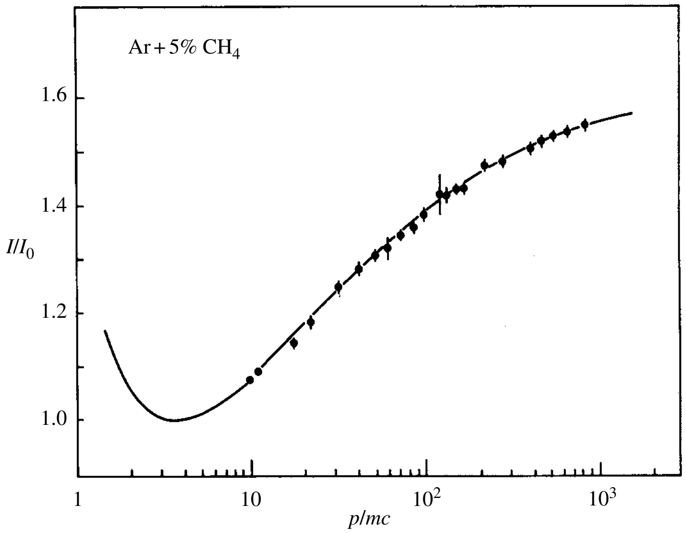


Fig. 9.8 Mean ionization energy loss of charged particles in an argon–methane gas mixture, plotted as a function of momentum in mass units, p/mc. The measurements were made by multiple ionization sampling, and show the relativistic rise from the minimum to a plateau. (After Lehraus *et al.* 1978.)

is independent of the mass M of the incident particle. It varies as $1/v^2$ at low velocity. After passing through a minimum value at an energy of about $3Mc^2$, the ionization loss increases logarithmically with energy. At higher energies, polarization effects set in and the ionization loss reaches a *plateau value* of about 2 MeV $\text{gm}^{-1}\,\text{cm}^2$, as in Fig. 9.8. Note also that, since $Z/A \sim 1/2$ in most materials (except hydrogen and the very heavy elements) the energy loss, expressed per gm cm^{-2} of material traversed, depends little on the medium.

9.6.2 Coulomb scattering

In traversing a medium, a charged particle will undergo electromagnetic (Coulomb) scattering by the nuclei of the medium (scattering by atomic electrons, because of their much smaller mass, is negligible by comparison). Individual scatters in angle are described by the Rutherford formula (1.23). Over a finite path length t, successive scatters combine to form a *multiple Coulomb scattering* distribution, which is approximately Gaussian with a root mean square value

$$\varphi_{\text{rms}} = \left(\frac{zE_{\text{s}}}{pv}\right)\sqrt{\frac{t}{X_0}} \tag{9.16}$$

where $E_s = (4\pi/\alpha)^{1/2}\ mc^2 = 21\,\text{MeV}$ and X_0 is the *radiation length* given by

$$\frac{1}{X_0} = 4\alpha\left(\frac{Z}{A}\right)(Z+1)\,r_{\text{e}}^2 N_0 \ln\left(\frac{183}{Z^{1/3}}\right) \tag{9.17}$$

Here r_{e} is the classical radius of the electron and $\alpha = 1/137$. Thus a singly charged particle with a value of $pv = p\beta c$ measured in MeV will suffer an rms deflection of $21/(pv)$ radians in traversing one radiation length.

9.6.3 Radiation loss

In addition to undergoing ionization energy loss and Coulomb scattering, high-energy electrons also suffer *radiation loss* with the emission of photons, a process known as *'bremsstrahlung'* or braking radiation. The average rate of such radiative energy loss of an electron in traversing a thickness dx of medium is

$$\left\langle \frac{\mathrm{d}E}{\mathrm{d}x} \right\rangle_{\mathrm{rad}} = -\frac{E}{X_0} \tag{9.18}$$

As the radiation probability is proportional to the square of the acceleration, $X_0 \propto 1/r_{\mathrm{e}}^2 \propto m_e^2$ and the radiation length for a muon will be of order $(m_\mu/m_e)^2$ times that for an electron. So in most practical cases, only radiation losses by electrons need be considered (radiation by muons is only significant for extremely high-energy muons capable of penetrating deep underground through km thicknesses of rock). If ionization losses are neglected, the mean energy of an electron of initial energy E_0 after having traversed a thickness x of medium will then be

$$\langle E \rangle = E_0 \exp\left(-\frac{x}{X_0}\right) \tag{9.19}$$

We see from (9.17) that the radiation length varies approximately as $1/Z$. For example, it is 40 g cm^{-2} in air, compared with 6 gm cm^{-2} in lead. While the rate of ionization energy loss is practically constant for high-energy electrons, the rate of radiation loss is proportional to energy E. The energy at which $(\mathrm{d}E/\mathrm{d}x)_{\mathrm{ion}} = (\mathrm{d}E/\mathrm{d}x)_{\mathrm{rad}}$ is known as the *critical energy* E_{c}. Electrons above the critical energy lose their energy principally through radiation processes, while for those below the critical energy it is mostly through ionization. Roughly, $E_{\mathrm{c}} \sim 600/Z$ MeV.

9.6.4 Pair production by γ-rays

Provided they have energy $E_\gamma > 2\, mc^2$, the photons radiated by an electron can themselves transform into e$^+$e$^-$ pairs, again in the Coulomb field of a nucleus to conserve momentum. The mean distance travelled by a photon before converting to a pair in a medium is called the *conversion length*. The conversion length is energy dependent, but at high energies (GeV) has an asymptotic value of approximately $(9/7)\,X_0$.

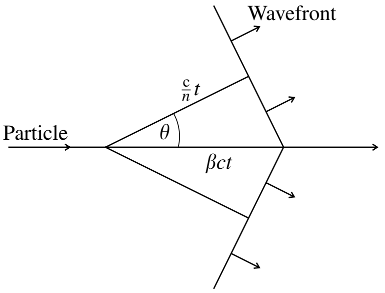


Fig. 9.9 Huyghens construction for emission of Cerenkov light by a relativistic particle.

9.6.5 Cerenkov radiation

As relativistic particles traverse a medium (e.g. the atmosphere), a small part of the energy loss appears in the form of a coherent wavefront of *Cerenkov radiation* (somewhat akin to the bow wave at the stem of a ship) as shown in Fig. 9.9. This radiation is mostly in the ultraviolet or blue region of the spectrum. The Huyghens construction in the figure gives

$$\cos\theta = \frac{(ct/n)}{\beta ct} = \frac{1}{\beta n}, \quad \beta > \frac{1}{n} \tag{9.20}$$

where the refractive index n of the air at ground level is given by $\varepsilon = n - 1 = 3 \times 10^{-4}$, a quantity which is proportional to air pressure. The threshold energy for an electron with $mc^2 = 0.51\,\mathrm{MeV}$ will be $mc^2/(1-\beta^2)^{1/2} = mc^2/(2\varepsilon)^{1/2} = 21$ MeV, while for a muon of $mc^2 = 106\,\mathrm{MeV}$ it is 4.3 GeV. Most of the components of air showers have much greater energies, so they will produce abundant Cerenkov light. Typically, a relativistic particle above threshold generates about 10,000 photons per km of path near ground level (and less at high altitudes where the pressure is lower). This light can be detected by means of large spherical mirror arrays which direct it on to photomultipliers placed at the focus (see Fig. 9.11).

9.6.6 Atmospheric fluorescence

Charged particles traversing the atmosphere not only ionize, but also excite the atoms. Some of this appears in the form of *fluorescence* from nitrogen molecules, with typically 5000 photons per km of track length, again in the blue wavelength region (300–450 nm).This fluorescent light is emitted isotropically. On the other hand, Cerenkov light from atmospheric traversal by extreme relativistic particles is emitted in a narrow cone of angle $\theta \sim (2\varepsilon)^{1/2}(= 1.4^\circ$ at ground level, although Coulomb scattering of the electrons will considerably broaden this). This distinction is important from the point of view of shower detection, as described below.

9.7 Development of an electromagnetic cascade

We now discuss the longitudinal development of an electromagnetic shower in very simplified terms. Consider an electron of initial energy E_0 traversing a medium, neglecting ionization losses. In the first radiation length, suppose the electron radiates one photon, of energy $E_0/2$. In the next radiation length, suppose the photon converts to an electron–positron pair, each with energy $E_0/4$, and that the original electron radiates a further photon, also of energy $E_0/4$. Thus, after two radiation lengths, we have one photon, two electrons, and one positron, each of them with the energy $E_0/4$. Proceeding in this way, it follows that after t radiation lengths, we shall have electrons, positrons, and photons in approximately equal numbers, each with energy $E(t) = E_0/2^t$. We assume that this cascade multiplication process continues until the particle energy falls to $E = E_c$, the critical energy, when we suppose that ionization loss suddenly becomes dominant and that no further radiation or pair conversion processes are possible. Thus the cascade reaches a maximum and then ceases abruptly. The main features of this simple model are the following:

- The shower maximum is at a depth $t = t\,(\max) = \ln(E_0/E_c)/\ln 2$, that is, it increases logarithmically with primary energy E_0.
- The number of particles at shower maximum is $N\,(\max) = 2^{t(\max)} = E_0/E_c$, that is, proportional to the primary energy.
- The number of shower particles above energy E is equal to the number created at depths less than $t(E)$, that is,

$$N\,(> E) = \int 2^t \mathrm{d}t = \int \exp\,(t\,\ln 2)\,\mathrm{d}t = \frac{(E_0/E)}{\ln 2}.$$

so that the differential energy spectrum of the particles is

$$\frac{\mathrm{d}N}{\mathrm{d}E} \propto \frac{\mathrm{d}E}{E^2}.$$

- The total track-length integral (of charged particles of $E > E_c$) in radiation lengths is

$$L = \left(\frac{2}{3}\right) \int 2^t \mathrm{d}t \sim \left(\frac{2}{3} \ln 2\right) \frac{E_0}{E_\mathrm{c}} \sim \frac{E_0}{E_\mathrm{c}}$$

This last result also follows from energy conservation: since the ionization loss per particle is E_c per radiation length, essentially all the incident energy is finally dissipated as ionization energy loss. Thus we obtain the very important result that the track-length integral gives a measurement of the primary energy.

Example 9.3 *A photon of energy 10 TeV is incident vertically on the atmosphere. Estimate the height of the maximum of the ensuing electron–photon shower in km. The critical energy in air is 100 MeV and the radiation length is 37 g cm^{-2}.*

From the above simple model, the depth of the maximum is $x = \ln(E_0/E_\mathrm{c})/\ln 2 = 16.6$ radiation lengths or 615 g cm^{-2}. Using the expression for the exponential atmosphere (9.14) results in a height for the maximum of 3.4 km.

In practice, of course, the effects of both radiation and ionization losses are present throughout the shower process, and an actual shower consists of an initial exponential rise, a broad maximum, and a gradual decline thereafter, as shown in Fig. 9.10. Nevertheless, the above simple model reproduces many of the essential quantitative features of actual electromagnetic cascades.

Our model has treated the shower as one dimensional. Actual showers spread out laterally, due mostly to Coulomb scattering of the electrons as they traverse the medium. The *lateral spread* of a shower of initial energy E_0 (in MeV) is a few times the so-called *Moliere unit*, equal to $E_\mathrm{s}/E_0 = 21/E_0$ radiation lengths.

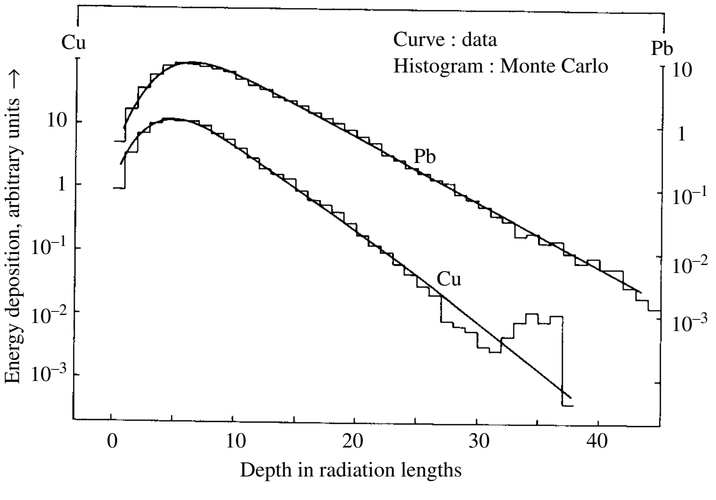


Fig. 9.10 Longitudinal development of electromagnetic showers due to 6 GeV electrons in CERN experiments (after Bathow *et al.* 1970). It is left as an exercise for the reader to show that the observed shower maximum occurs rather earlier in the cascade than the above simple model suggests.

9.8 Extensive air showers: nucleon- and photon-induced showers

If the primary particle is a high-energy proton or heavier nucleus rather than an electron or photon, a *nuclear cascade* will develop through the atmosphere. The longitudinal scale is the nuclear interaction length in air, $\lambda_{\text{int}} \sim 100\ \text{gm cm}^{-2}$. The proton (or heavier nucleus) generates mesons in these interactions, and they can in turn generate further particles in subsequent collisions. While, in the electron–photon shower, the electrons lose the bulk of their energies in a radiation length, the nucleons can in general penetrate through several interaction lengths, losing only a fraction of their energy—typically 25%—in each encounter, to struck nucleons as well as mesons. In air, the nuclear interaction length is some 2.5 times the radiation length $X_0 \sim 40\,\text{gm cm}^{-2}$. Thus cascades initiated by nucleons are very much more penetrating than purely electromagnetic cascades initiated by photons, and this difference in the shower profile through the atmosphere has been exploited to differentiate between nucleon- and photon-induced cascades.

Another difference is that the lateral spread of nuclear showers is determined mostly by the transverse momentum of the secondaries in nuclear interactions, typically 0.3 GeV/c, and is much larger than for an electromagnetic shower of the same primary energy. Such *extensive air showers* will contain a high-energy core, predominantly of nucleons, with a more widely spread electron–photon component which is continually fed by fresh neutral pion production and decay, $\pi^0 \to 2\gamma$, and fresh electromagnetic cascades. As mentioned in Section 9.6, the pion decay probability at high energy falls off as $100/E_\pi\,(\text{GeV})$, so that apart from a small amount of energy in the form of neutrinos and muons from pion decay, the great bulk of the energy as well as the overwhelming majority of the particles in a proton-initiated extensive air shower will end up in electron–photon cascades. So the track-length integral will again give a measure of the primary energy of the shower. However, the interpretation of the actual signals recorded by either ground arrays or air Cerenkov/fluorescence detectors does depend to some extent on the modelling used to simulate the nucleon cascade, and it is noteworthy that the inferred fluxes from the AGASA, HiRes, and AUGER arrays differ typically by a factor of up to 2.

9.9 Detection of extensive air showers

The detection of extensive air showers has been accomplished by a variety of techniques. The oldest technique, pioneered by Auger more than 75 years ago, uses an extended *array of ground detectors* in coincidence. These ground detection detectors sample the charged particles in the shower, usually by means of tanks containing liquid scintillator or water Cerenkov counters. Such showers only become detectable at sea-level for primary energies $E_0 > 1000\,\text{TeV}$ (10^{15}eV), when the maximum occurs near the ground. At mountain altitudes, the threshold is typically 100 TeV. The particles in such showers all have $v \sim c$, so that the shower front is quite well defined, and the direction of the primary particle can be measured rather accurately by timing the different parts of the shower front as it crosses the array.

The other established technique for shower detection is the use of mirror plus photomultiplier systems to record the Cerenkov light and/or fluorescent light from the atmosphere itself. As explained above, atmospheric Cerenkov light appears in a narrow angular range, so that the light pool appears in a restricted radius of order 100 m around the shower axis, which must be fairly close to the mirror system to record any signal. On the contrary, the fluorescent output is isotropic, so that the light pool is much more extensive, and this means that distant showers several kilometres away, not aimed towards the mirror/photomultiplier system, can be detected, and therefore the sensitivity to the highest energy—and rarest—events is greatly increased. The weakness of these techniques is that they have a poor duty cycle. The problem of stray background light can only be overcome by operating on cloudless, moonless nights. Even in the most favourable environments, the duty cycle is only 10%.

The first system exploiting the 'atmospheric light' technique was the mirror array at the Whipple Observatory (see Fig. 9.11). It employs two spatially separated arrays of mirrors to give stereo images of the showers from the

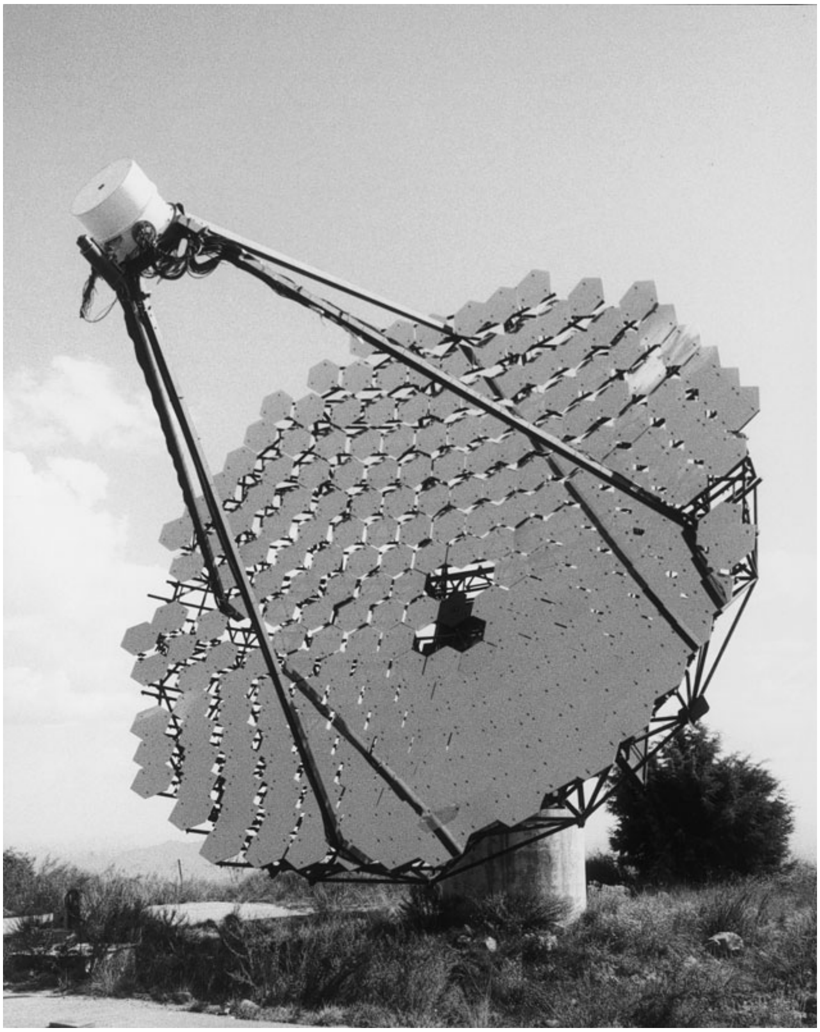


Fig. 9.11 Photograph of the 10 m mirror array at the Whipple Observatory, Arizona, for detection of Cerenkov light from air showers. A second nearby (11 m) dish allows construction of stereo images of extensive air shower profiles. These have recently been replaced by the larger, four-dish VERITAS array. The Whipple Observatory was the first to identify a point source of very high-energy γ-rays (>1 TeV) from the Crab Nebula. (Photo courtesy of Trevor Weekes 1998.).

light output. In this way, it is possible to reconstruct the shower profile and to distinguish showers initiated by primary photons, which develop early and are contained in the upper atmosphere, from those generated by nucleons, which develop more slowly and are more penetrating. This feature is valuable in identifying point sources of γ-rays.

9.10 Point sources of γ-rays

The majority of cosmic γ-rays form a steady, random background of secondary origin, coming, for example, from the decay of neutral pions produced when primary protons interact with interstellar matter. Nevertheless, using instruments such as EGRET (Energetic Gamma Ray Experimental Telescope), on the GRO (Gamma Ray Observatory) satellite launched in 1991 (see Fig. 9.12), point sources have been detected in the range of quantum energy from 3 keV to 30 GeV; and using the ground-based air Cerenkov method (Fig. 9.11) to energies of 10 TeV.

Many known pulsar sources (see Section 10.10) such as Crab, Geminga, and Vela have been detected in this way by different laboratories (see examples in Fig. 9.13). The main mechanism producing the γ-rays is believed to be that of radiation by electrons in the intense magnetic fields of the pulsars, which is in the form of synchrotron radiation.

For a given electron energy, the energy spectrum of the radiated photons is roughly of the form $\mathrm{d}E/E$, that is, peaked to low energies. However, the electron and photon intensities near such sources are so high that very energetic photons can be produced via inverse Compton effect, that is, low-energy photons being boosted by collisions with the very energetic electrons (which in turn are the products of shock acceleration).

The sources described here are steady sources of γ-rays. The source in the Crab, for example, originated nearly one thousand years ago (the A.D.1054 supernova), and is a sort of 'standard candle' for γ-ray spectroscopy. Figure 9.14 shows the Crab spectrum from the MAGIC array, which is a 17 m diameter single dish in the Canary Islands using air Cerenkov detection. The spectrum

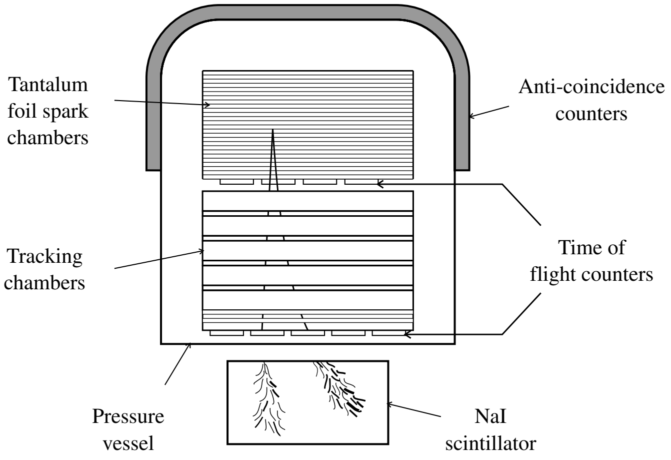


Fig. 9.12 Diagram of the EGRET instrument on the GRO satellite. γ-Rays are detected when they materialize into e^+e^- pairs in the upper stack of tantalum sheets and spark chambers, the total energy of the ensuing electron–photon shower being measured from pulse heights in the lower array of NaI scintillators (after Ong 1998).

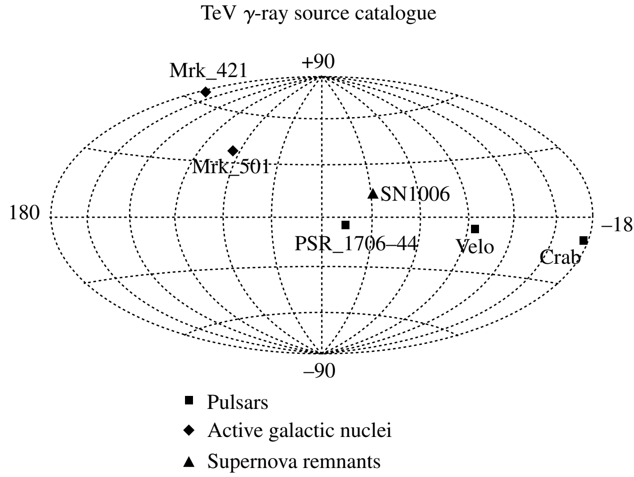


Fig. 9.13 Map of γ-ray point sources detected by the EGRET satellite experiment, for gamma energies above 100 MeV. The coordinates are the longitude and latitude relative to the plane of our galaxy (from Ong 1998).

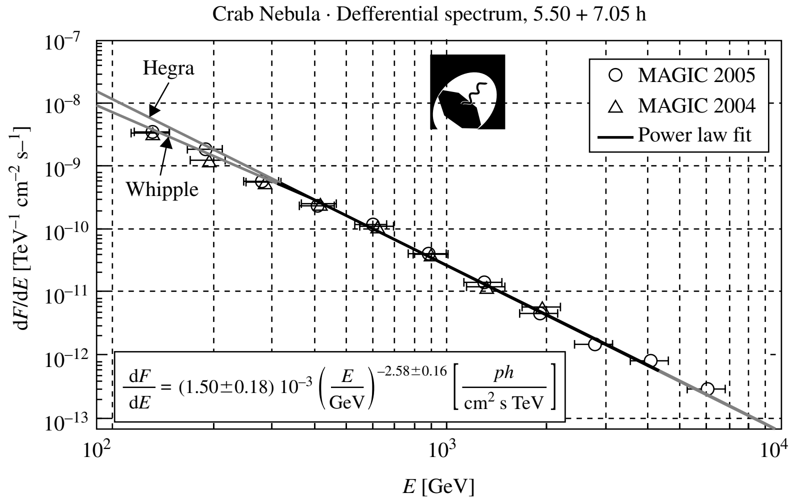


Fig. 9.14 γ-Ray spectrum from the Crab nebula, measured by the MAGIC air Cerenkov array (from Wagner *et al.* 2005).

(Wagner *et al.* 2005) up to 10 TeV energy has the form for the flux of γ-rays m^{-2} s^{-1} TeV^{-1}

$$N\,(E)\,\mathrm{d}E = \frac{(24 \pm 3)\cdot 10^{-8}}{E^{2.6\pm0.2}}\mathrm{d}E \tag{9.21}$$

with E in TeV. Over a period, the flux has been measured to be constant within 1% accuracy.

Some γ-ray sources have been identified with the AGNs described below, with redshifts up to $z \sim 2.5$. These sources do vary with time. They, just like the quasars –with which they may even be identified—are considered to be associated with very massive black holes located at galactic centres, because it seems that black holes are the only compact sources capable of generating such enormous energies and intensities of radiation, principally concentrated in the 'jets' described in the radio galaxies in Section 9.14. Presumably, like the radio emission, the γ-rays originate from these jets as electrons spiral in the magnetic field of the jet and emit synchrotron radiation.

9.11 γ-Ray bursts

More dramatic than the relatively steady γ-ray sources described above, are the *γ-ray bursts*, which last typically from 10 milliseconds to 10 seconds. The quantum energy of this radiation is in the region of 0.1–100 MeV, with the shorter bursts associated with harder spectra. Such bursts were first observed by accident in the 1960s, by the US Vela satellite, searching for radiation from underground nuclear tests. The events are quite common—around one or two per day—and the sources seem to be distributed more or less isotropically over the sky, as shown in the all-sky map of Fig. 9.15. Since the γ rays have a continuous, rather than line spectrum, the redshift was initially unknown, until their extra-galactic origin was finally established in the 1990s, when a burst was pinpointed by the Dutch–Italian BeppoSAX satellite and the optical afterglow was found to have a spectrum indicating a redshift of $z \sim 0.8$. Most of the bursts have $z > 1$, the present record being $z = 6.4$, corresponding to an event taking place when the universe was only about one billion years old.

The amount of energy in these bursts is $\sim 10^{44}$ J, comparable with or even greater than that emitted (in photons) in the course of a Type II supernova explosion (the total energy release in this case is 10^{46} J, but 99% of that is in the form of neutrinos—see Section 10.9). The bursts can be grouped into two classes, with apparently different origins; short-duration bursts, with a period less than 2 s, and an average of 0.3 s; and long-duration bursts, with a period from 2 to 10 s.

It seems that the *long-duration bursts* are becoming better understood. They come from events called *collapsars*, for which the progenitors are a class of very massive (20–100 solar mass) stars identified with the rapidly rotating, low-metallicity Wolf–Rayet (W–R) stars. It is believed that, at an early stage in their evolution, galaxies were dominated by such massive stars, which would have evolved very rapidly, since as shown in Chapter 10, the lifetime of a star of mass M on the main sequence varies as $M^{-2.5}$. Thus a star of 100 solar masses would have a lifetime some 10^{-5} of that of the Sun, or less than 100,000 years. Formed at an early stage in galactic evolution, the abundance of heavy elements in such stars would be low (in distinction to stars forming in later stages, out of previous generations of recycled stellar ejecta). Because of their large mass,

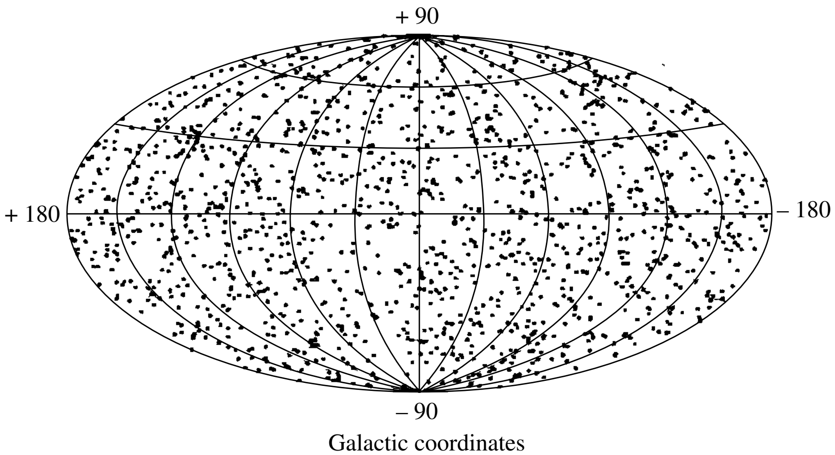


Fig. 9.15 All-sky map of γ-ray bursts, indicating uniform distribution throughout the sky, from the burst and transient source experiment (BATSE) flown in the GRO satellite.

the W–R stars evolve rapidly, and at the end of silicon burning, the core of such stars is believed to collapse directly to a black hole—as distinct from stars of lower mass—passing first through the neutron star stage. The blast of γ-rays accompanies the enormous energy released when the surrounding material in the accretion disc is sucked into the black hole.

We should note here that, as will be explained in Chapter 10, stars form from accumulating interstellar gas falling into orbital motion about the centre of mass, and this so-called protostar evolves through successive stages of thermonuclear burning, with increasing temperature and enormous contraction of the core. Because of contraction and conservation of angular momentum, the core will generally be rapidly rotating. As in the case of quasars, because of the torus shape of the spinning material, the γ-rays which escape seem to be confined to narrow jets—typically with 3° opening angle—along the rotation axis. Of course, only those bursts directed towards the Earth will be observed, so that the actual rate integrated over all directions and over the whole sky could be as high as one per minute. A sub-class of collapsars are the so-called soft gamma repeaters, in which the bursts come from a single source at irregular intervals, and contain very soft γ-rays or X-rays.

The *short-duration bursts* are not well understood. How they originate is presently unclear; however, the very short burst lengths indicate an origin from extremely compact objects. They might arise, for example, from orbiting neutron stars which are components of a binary system, as they collapse under energy loss due to gravitational radiation to form black holes. The rate is about 10,000 times less than that of supernova explosions resulting in neutron stars, so that the overall observed rate of about 1 burst per day is compatible with the estimated rate of neutron star mergers.

In summary, γ-ray bursts are some of the most energetic, most intriguing, and most perplexing events in the cosmos. A big research effort is currently under way, and in time the basic mechanisms underlying these events will become much better understood.

Example 9.4 *High-energy γ-rays from very distant sources may encounter a cut-off in energy due to collisions with photons of the cosmic microwave background, or of starlight (optical or infrared), through formation of electron–positron pairs. Estimate the threshold energies involved, and the relevant absorption lengths.*

The absorption length for high-energy γ-rays is determined by the processes involved. From (1.26), Compton scattering $\gamma e \to \gamma e$ and pair production $\gamma\gamma \to \mathrm{e}^+\mathrm{e}^-$ well above threshold have rather similar cross-sections. However, since the CMB photon number density in intergalactic space is many orders of magnitude larger than that of the electrons, above that threshold, pair production off CMB photons will totally dominate.

If E_{th} denotes the threshold photon energy, E_0 the quantum energy of the target photon, the condition for pair production, $\gamma\gamma \to \mathrm{e}^+\mathrm{e}^-$ is $s = (E_{\text{th}} + E_0)^2 - (\mathbf{p_{th}} + \mathbf{p_o})^2 > 4m^2$, where m is the electron mass. For a head-on collision, $E_{\text{th}} = m^2/E_0$. For microwave photons $T = 2.73\,\text{K}$, $kT = 2.35 \times 10^{-4}\,\text{eV}$, and for one of energy $E_0 = y \cdot (kT)$, the value of $E_{\text{th}} \sim 10^3/y$ TeV. Most of the collisions are not head on, but we can take this as a typical threshold energy.

The mean free path for creating electron–positron pairs is $\lambda = 1/(\rho\sigma)$ where ρ is the density of microwave target photons and σ is the pair-production cross-section. At the threshold $s_{\text{th}} = 4m^2$, the cross-section is zero, and it rises with energy to a maximum of $\sim 0.25\sigma_{\text{Thomson}}$ at $s \sim 8m^2$ (see (1.26)), before falling off at higher energies. We take $\sigma \sim \pi\alpha^2/m^2 = 2.5 \times 10^{-25}\,\text{cm}^2$ as indicative of an upper limit. Inserting $\rho \sim 400\,\text{cm}^{-3}$ for the total microwave photon number density results in $\lambda \sim 10^{22}\,\text{cm} \sim 4\,\text{kpc}$. Although this is only a rough lower limit, it shows that for photons of energies above 10^3 TeV (10^{15} eV), the universe on the scale of megaparsecs will be quite opaque. On the other hand, less than 10^{-6} of the microwave photons have $y > 20$ in the high-energy tail of the Planck spectrum, and for these the threshold energy is only 50 TeV while the mean free path is over 100 Mpc, and such photons could be received from point sources spread over a considerable fraction of the universe (one must remember here that at high redshifts, the microwave photon energies, and hence threshold energies, are increased by a $(1 + z)$ factor).

For starlight photons, we can take the solar photosphere temperature of 6000 K, that is, a quantum energy of 6000/2.73 times that of the microwave photons, with a correspondingly reduced threshold energy of typically 1 TeV. The galactic starlight energy density is of order 1 eV cm^{-3} or $\rho \sim 1$ photons cm^{-3}, hence $\lambda > 1$ Mpc, large compared with the galactic radius. So there would be no problem in detecting γ-rays of any energy from sources in the local galaxy.

9.12 Ultra-high-energy cosmic ray showers: the GZK cut-off

As shown in Fig. 9.3, the spectrum of charged primary cosmic rays extends to at least 10^{20} eV. The data at these energies comes from very extensive counter arrays (at sea-level or on mountains). At present, the largest arrays probing the highest energies are AGASA in Japan, covering 100 km^2 with surface detectors; the HiRes experiment in Utah, detecting fluorescence radiation only; and the AUGER experiment in Argentina, covering 3000 km^2 with a surface detector array and also equipped with fluorescence detectors.

The HiRes project (Abassi *et al.* 2005, 2007) is in the form of a ‘fly’s eye’, namely an array of 67 1.6 m diameter hemispherical mirrors, each equipped with 12 or 14 photomultipliers (PMTs) at the focus. The mirrors are oriented so that together they cover the entire sky, each of the total number of 880 PMTs covering a 5° × 5° pixel.

The AUGER experiment (Abraham *et al.* 2004) is instrumented with a surface array of 1600 water tanks to record the Cerenkov light emitted as relativistic shower particles traverse the water, as well as a set of 240 detectors in four stations to record the fluorescent light emitted by nitrogen molecules excited by the shower particles as they traverse the atmosphere (see Fig. 9.16).

As already indicated in Fig. 9.3, the spectrum shows a ‘knee’ followed by an increasing slope of the spectrum above 10^{15} eV, and an ‘ankle’ at 4×10^{18} eV indicating a flattening of the spectrum, followed by a catastrophic fall-off above

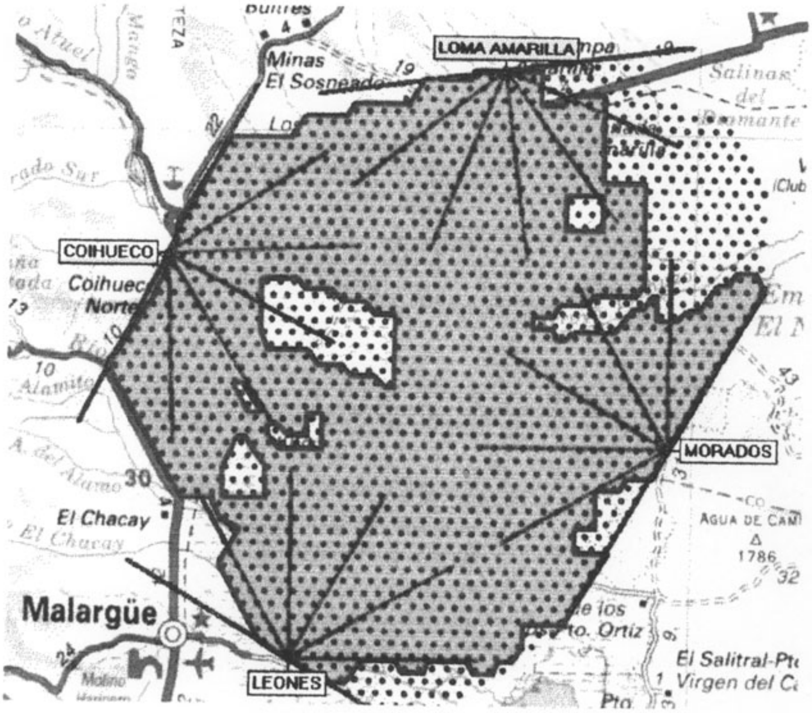


Fig. 9.16 Map of the AUGER extensive air shower array. The 1600 water Cerenkov tanks forming the ground detector array are shown as dots, at 1.5 km separation. They are overlooked by four stations housing 240 mirror/photomultiplier arrays, which record the fluorescence from nitrogen molecules excited as the air shower traverses the atmosphere. The ability to combine the data from the ground array with that from air fluorescence has proved a powerful constraint on energy measurements. (courtesy Pierre Auger collaboration).

4×10^{19} eV. Many years ago, Greisen (1966), Zatsepin and Kuzmin (1966) pointed out that the universe could become opaque at such energies through photopion production excited in collisions of the primary protons with photons of the microwave background radiation, known as the GZK effect:

$$\begin{aligned} \gamma + p \to \Delta^+ &\to p + \pi^0 \\ &\to n + \pi^+ \end{aligned} \tag{9.22}$$

If the proton has mass M, momentum **p,** and energy E, and the microwave photon has momentum **q** and energy $qc,$ then the square of the total centre-of-momentum energy in the collision will be (see Chapter 2 on relativistic kinematics):

$$\begin{aligned} s = E^2_{\text{cms}} &= (E + q)^2 - (\mathbf{p} + \mathbf{q})^2 \\ &= M^2 + 2q\,(E - |\mathbf{p}|\cos\theta) \end{aligned}$$

in units $c = 1$. Here θ is the angle between the proton and photon directions, and E_{cms} must be at least equal to the sum of proton and pion masses. The threshold proton energy then becomes

$$\mathrm{E}_{\text{th}} = \frac{5.96 \times 10^{20}}{[\mathrm{y}(1 - \cos\theta)]}\mathrm{eV} \tag{9.23}$$

where we take the photon energy as $q = y\,kT$ and the microwave background has $kT = 2.35 \times 10^{-4}$ eV. Typically, a proton would lose 15% of its energy in such a collision.

As an example, for head-on collisions ($\cos\theta = -1$) and $y = 5, E_{\text{th}} = 6 \times 10^{19}$ eV. The cross-section for the reaction (9.22) near threshold is $\sigma = 2 \times 10^{-28}\,\text{cm}^2$, and the total microwave photon density is $\rho = 400\,\text{cm}^{-3}$, giving a collision mean free path $\lambda = 1/\rho\sigma = 4.1$ Mpc for all the microwave photons. For the 10% of photons with $y > 5$ the mean free path would be of order 50 Mpc. So one can expect that protons of $E > E_{\text{th}}$ coming from beyond the local galactic supercluster would have their energies attenuated by collisions with the microwave background (see Table 5.1 for distance scales). In fact detailed calculations (Dermer 2007) show that indeed the flux above about 5×10^{19} eV should fall off sharply, as seems to be found in the HiRes and AUGER experiments—see (9.2) and Fig. 9.3(c). The AUGER results on shower profiles indicate that many of the highest energy primary particles are heavy nuclei, for which the dominant energy loss mechanism would be photonuclear disintegration by CMB photons. It turns out however that the expected cut-off energy is comparable to that for the protons.

It is believed that the 'ankle' effect, that is, the slight flattening of the spectrum between 4×10^{18} eV and 4×10^{19} eV, may indicate the dominance of particles which are extra-galactic in origin. For a galactic magnetic field of order 3μ G (0.3 nT) the radius of curvature of protons of $E = 4 \times 10^{18}$ eV is ~ 4 kpc. Such particles would not be trapped in the magnetic fields of a spiral galaxy, with a disc thickness an order of magnitude less than this.

9.13 Point sources of ultra-high-energy cosmic rays

There is in general no correlation between the directions of high-energy (charged) cosmic ray primaries and point sources. Except at the very highest energies, any anisotropy would be destroyed by the galactic magnetic field. However, the AUGER experiment does find significant anisotropies for events above 6×10^{19} eV (Abraham *et al.* 2007). All 27 events above this limit are found to lie inside a 3° cone towards a known AGN (located within 100 Mpc of the Earth). The probability that this could happen by chance is of order 2×10^{-3}. This evidence seems to indicate that the extreme energy cosmic rays are accelerated by a mechanism associated with massive black holes. Of course, the fact that such correlations are possible implies that the net magnetic fields in deep space beyond 10 Mpc must be extremely weak, less than about 10^{-11} G (see also Section 8.2).

9.14 Radio galaxies and quasars

The electromagnetic radiation from a galaxy such as the Milky Way encompasses a vast spectral range, from radio wavelengths (centimetres to kilometres) to γ-rays with energies up to many TeV. In our own galaxy, less than

1% of the total electromagnetic output is at radio wavelengths, but so-called radio galaxies are observed in which the radio emission can far exceed the optical output (from stars). The most dramatic radio emission is from *quasars* (standing for quasi-stellar radio sources), which are the brightest optical and radio sources in the sky, far exceeding the total light output from their host galaxies. For this reason they can be detected even at enormous distances. In fact, quasars are almost invariably found to have large redshifts; more than half have $z > 1$ and the largest one seen to date has $z = 6.4$. Indeed, it was these very large redshifts of the spectra that made the original identification of quasars as highly luminous but very distant objects so difficult. Quasars in fact correspond to the most distant events known, occurring at times $t \sim t_0/(1+z)^{3/2}$ in the development of the universe. Nearly 60,000 quasars have been observed to date. They are indeed to some extent an ancient phenomenon, typically occurring billions of years ago, at an early stage in the evolution of galaxies. We observe them today because of their great distance and the finiteness of the velocity of light or radio waves.

9.14.1 Radio telescopes

Quasars are often associated with galaxies so distant that the optical signal is hardly detectable, and their original discovery was made with radio telescopes which incorporate giant receiving dishes. The largest single dish is the fixed one at Arecibo (see Fig. 9.17). Radio telescopes can have several advantages over optical telescopes. The radio signal does not suffer appreciable absorption by gas and dust, so one can probe deep into galactic centres. The signal can be amplified electronically; and its phase can be measured, so that the amplitudes of signals from several separate telescopes on a very long baseline L can be combined coherently, using atomic clock timing signals and optical fibre transmission to a central analysis receiver. This procedure yields an effective aperture of L and a very high angular resolution $\Delta\theta = \lambda/L$ comparable to the best optical telescopes (see Fig. 9.18).

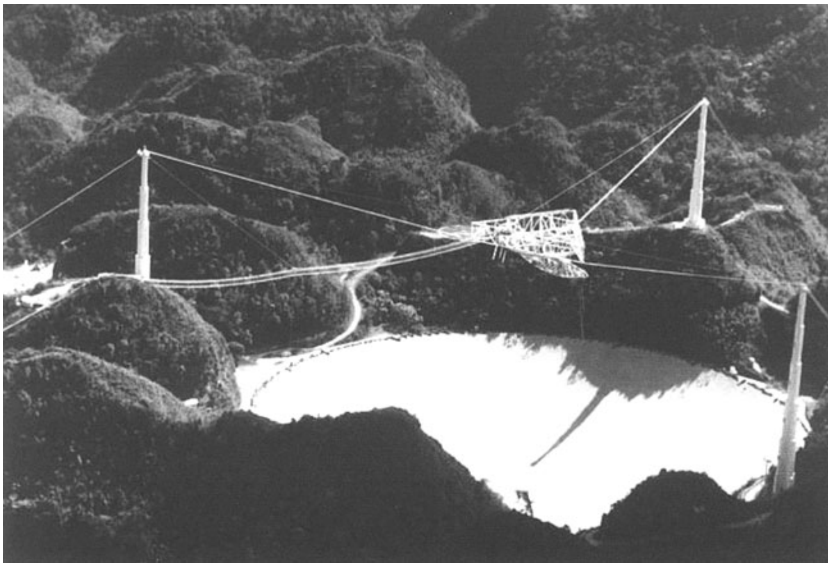


Fig. 9.17 The Arecibo radio telescope in Puerto Rico is the world's largest single-dish instrument, with a diameter of 300 m, built into a natural depression. The receiver (or transmitter) is suspended at the focus of the dish from three pylons. In its 40-year history, it was responsible for the discoveries of the first binary pulsar and the first extra-solar planets. (Courtesy National Astronomy and Ionosphere Center, Cornell and NSF).

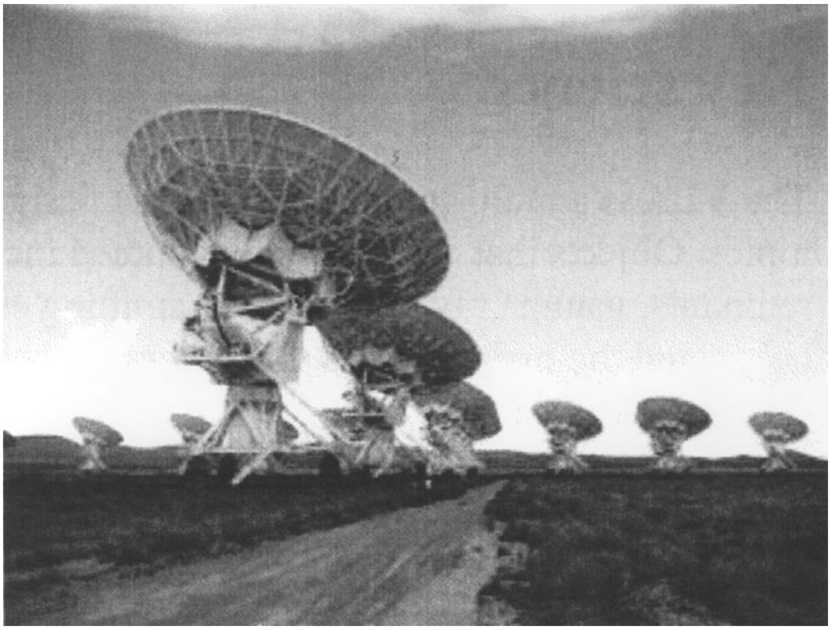


Fig. 9.18 The multiple dishes of the VLA (very large array) radio telescope, operated by the National Radio Astronomy Observatory (NRAO) in New Mexico. It consists of twenty seven 25 m diameter dishes, which can be operated as an interferometer with a baseline L up to 36 km, and an angular resolution for 7 mm wavelength of $\lambda/L \sim 0.2\,\mu$rad or 0.05 arcsec.

9.14.2 Active galactic nuclei (AGNs)

Quasars are associated with so-called active galactic nuclei (AGNs). Only about 1% of all galaxies fall into this class. The quasar luminosities often vary with time, on a timescale of months or days, indicating a source of limited spatial extent (light-months or light-days). This fact, together with the enormous quasar luminosities, typically 10^{40} W, or about 10^{13} times that of the Sun (3.9×10^{26} W), leaves massive black holes as about the only conceivable objects which could provide such power in a compact region of space, by consuming inflowing matter (equivalent to roughly one solar mass energy per year). Quasars are believed to be associated with black holes of typically $10^6 - 10^9$ solar masses at galactic centres. Black holes have been mentioned in Chapter 2 and are further discussed in Chapter 10. They are objects with such strong gravitational fields that even relativistic particles such as photons are trapped inside them, and nearby material is attracted by the strong field, flows into the black hole and is gobbled up.

A massive black hole would be surrounded by a spinning pancake-like accretion disc of galactic material—gas, dust, and stars—which feeds its growth. The high rate of spin just results from the contraction and angular momentum conservation (rather like bathwater going down a plughole). The energy emitted, mainly in the infrared region of the spectrum, is provided by the gravitational energy released when the material in the accretion disc is swallowed up by the black hole. In this process, about half the mass energy of the material would be released, so that a diet of a few solar mass stars per year would be enough to provide the above energy output. However, this accretion cannot go on indefinitely. After the nearby material is exhausted, the AGN dies down, and one is left with a normal galaxy containing a massive but relatively quiescent black hole at the galactic centre—the ash, so to speak, of the AGN. For example, our own galaxy, the Milky Way has at its centre a black hole of 3.6 million solar masses, identified with the radio source Sagittarius A* (see Problem 10.8).

In the course of absorption, the accretion material will undergo violent oscillations and will be ionized as a plasma, with the result that charged particles can be accelerated to very high energies, and in the accompanying magnetic fields generated by the plasma currents, can radiate at infrared, optical, and X-ray frequencies. In some cases these charged particles can punch their way through the minor axis of the accretion disc, giving rise to two narrow jets of particles travelling in opposite directions. These jets create enormous lobes of plasma as they traverse the intergalactic medium, and it is this plasma which generates the radio emission, giving rise to the name 'radio galaxy'. Since magnetic fields will be associated with the jets of charged particles, the radio emission would be part of the *synchrotron radiation* produced. The name comes from particle accelerators called synchrotrons, in which electrons are confined and accelerated in circular paths by strong magnetic fields, and radiate quanta as a result of the acceleration.

Figure 9.19 shows a sketch of the two-jet process, and Fig. 9.20 a picture of a typical radio galaxy, Cygnus A. The jets in this case extend to several Mpc. Indeed, it has become clear that the nature of the phenomena associated with massive galactic black holes depends to a large extent on the angle between the

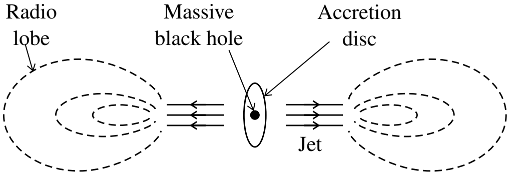


Fig. 9.19 Sketch of possible two-jet mechanism involved in radio emission from a quasar.

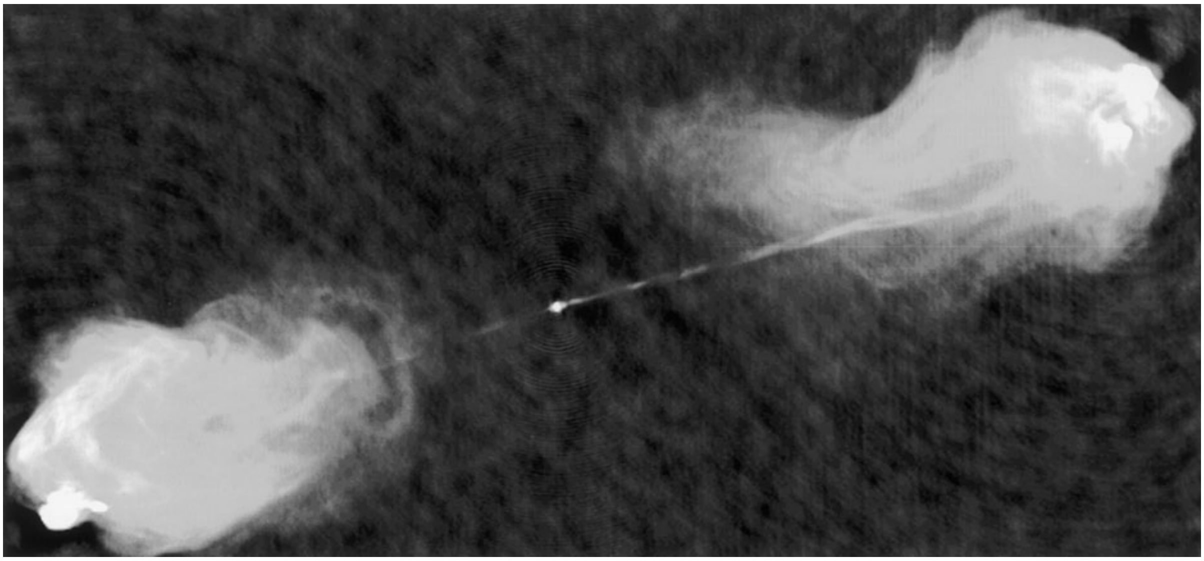


Fig. 9.20 Radio image of the radio galaxy Cygnus A. The galactic centre is the small dot midway between the massive radio lobes. (Courtesy Chris Carilli, NRAO, 2002).

jet axis and the line of sight to the Earth. If it is large, one obtains two jets of comparable size as in Fig. 9.20 in the case of Cygnus A. However, if the jet velocities are extreme relativistic and the angle happens to be small, so that one jet is approaching and the other is receding, the Doppler shift of the frequency can mean that the approaching jet is very bright while the receding one is below the threshold for detection. There are many interesting effects associated with the jets. For example, the observed transverse velocity may apparently exceed light velocity (see Problem 9.13). As another example, if the photons observed are of high energy, they would have been radiated by electrons of very large Lorentz factor γ, and hence are confined to a narrow angular spread of order $1/\gamma$ (see Problem 2.5). Thus small fluctuations in the jet angle could deflect the beam away from the observer, and the observed intensity could vary on short timescales. This may be a possible cause of the extreme variability of some of the γ-ray sources.

Two-jet phenomena, on a much smaller scale, and termed *microquasars* have also been observed in local galaxies. These are assumed to be due to black holes of order a few solar masses only, with accretion from a nearby companion star. The distance scale of the radio lobes is of order parsec in this case, rather than the megaparsec scale of quasars.

AGN sources from which the γ-ray emission is in the TeV energy region, are referred to as *blazars,* the most famous example being Markarian 501 (see Fig 9.13). From this source the γ-ray flux can vary by an order of magnitude from night to night, and during the year 1997 it increased by a factor 50.

9.14.3 The Lyman α forest

Before leaving the subject of quasars, we refer briefly to their role in establishing the degree of re-ionization of the stellar medium, consequent on the appearance of the first stars. As discussed in the next chapter, stars are believed to have started to form at redshift $z < 12$, and their ultraviolet emissions would have caused re-ionization of the, previously neutral, interstellar hydrogen. The degree of ionization can be estimated from observations of the so-called Lyman α forest. This arises from clouds of hydrogen when they are backlit by the intense light from a quasar. Let us recall that the Lyman series in the hydrogen atom consists of spectral lines of wavelength $\lambda = \left(2h/\alpha^2 m_e c\right)\left[(1/n_1)^2 - (1/n_2)^2\right]$ for transitions between principal quantum numbers $n_1 = 1$ and $n_2 > 1$. The Lyman α line, for $n_2 = 2$, has $\lambda = 121.6$ nm, while the 'Lyman limit', for $n_2 = \infty$, has $\lambda = 91.0$ nm. Passage of light from the quasar through any neutral hydrogen will show a 'forest' of absorption lines between these two wavelengths. Figure 9.21 shows a typical quasar spectrum (corrected for any redshift). From analysis of the line shapes and intensities it is then possible to deduce the fraction of hydrogen which is neutral, that is, not ionized. It is small but not negligible.

9.15 Atmospheric neutrinos: neutrino oscillations

Neutrinos and antineutrinos are constituents of the secondary cosmic rays generated in the Earth's atmosphere by interactions of the primary particles,

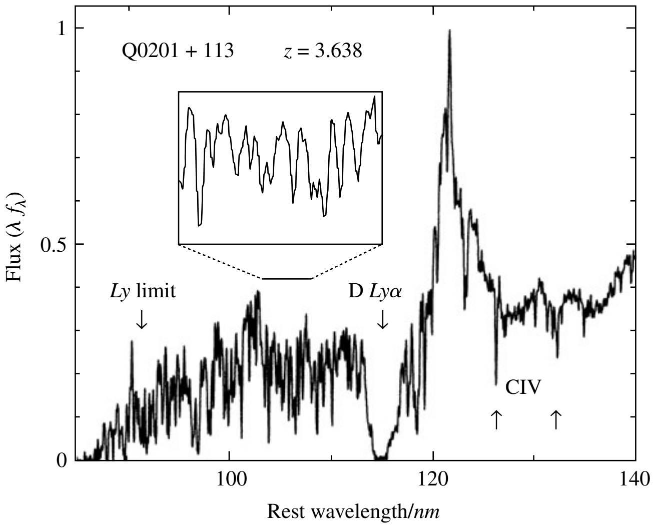


Fig. 9.21 The spectrum of hydrogen absorption lines, the so-called Lyman α forest produced by clouds of neutral hydrogen backlit by the ultraviolet light from a quasar (from 'Cosmological Physics' by J.A. Peacock 1999).

as discussed in Section 9.5 above. The primaries generate mesons (pions and kaons) in collisions with air nuclei. These in turn decay in flight into neutrinos and muons: for example, $\pi^+ \to \mu^+ + \nu_\mu$. The muons in turn decay to muon- and electron-neutrinos: for example, $\mu^+ \to e^+ + \nu_e + \bar{\nu}_\mu$. The neutrino energy spectrum peaks at around 0.25 GeV, and falls off as $E^{-2.7}$ at higher energy. At low energies, of order 1 GeV, both the pions and the muons will mostly undergo decay in flight (rather than interacting or coming to rest in the atmosphere), so that as the above decay modes suggest, the expected ratio of fluxes $\phi\left(\nu_\mu\right)/\phi\left(\nu_e\right) \approx 2$ in the GeV energy region, and this ratio should be reflected in the relative numbers of interactions with secondary muons or electrons (see Problem 9.7). Several underground experiments in the 1990s discovered on the contrary that the observed ratio of numbers of interactions containing muons as compared with electrons was of order 0.7 of that expected, signalling the possibility of a new phenomenon such as *neutrino flavour oscillations.*

Although the absolute flux of atmospheric neutrinos is low (about 1 cm^{-2}s^{-1} at sea-level) and the interaction cross-sections are feeble (of order 10^{-38} cm^2 per nucleon at a typical energy of 1 GeV), their interactions have been recorded in substantial numbers, starting in the late 1980s using large (multikiloton) underground detectors, originally intended to search for proton decay. The neutrino interactions were at first considered to be an annoying background incapable of eradication. In fact, a handful of atmospheric neutrino interactions had first been observed in small detectors placed deep underground in the early 1960s, but at that time they were considered to be of little interest. So the discovery of neutrino oscillations with atmospheric neutrinos, like many other discoveries in science, has been largely accidental, as a by-product of an investigation which failed in its original aim. By good fortune it turns out that the typical energies of atmospheric neutrinos, of order 1 GeV and determined by the effect of the Earth's magnetic field on the primary cosmic ray nuclei, combined with the accessible neutrino path lengths, determined by the Earth's radius, are exactly matched to the relevant scale of neutrino mass differences. As indicated in Chapter 4, there are three neutrino flavours (ν_e, ν_μ, and ν_τ) and

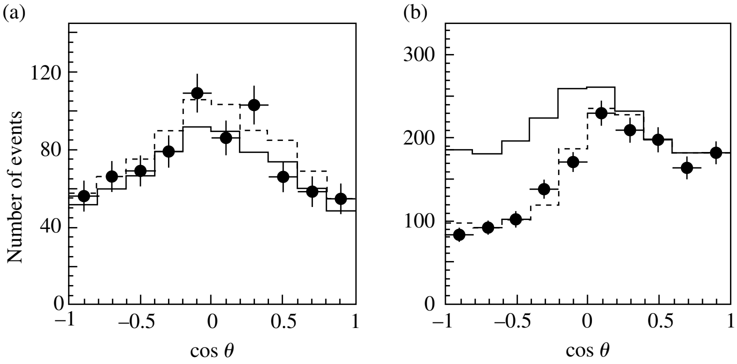


Fig. 9.22 Observed zenith angle distribution of (a) electron and (b) muon events with lepton momenta above 1.3 GeV/c in the Superkamiokande detector. The full-line histograms show the event rates expected for no oscillations, while the dashed histograms show the best-fit predictions for an oscillation scenario. The plots indicate no oscillations for electron-neutrinos, and for muon–neutrinos, maximum $\nu_\mu \to \nu_\tau$ mixing with $\Delta m^2 = 2.3 \times 10^{-3}\,\text{eV}^2$ (from Suzuki 2005).

three mass eigenstates (ν_1, ν_2, and ν_3), so there will be two independent mass differences. The larger one is associated with atmospheric neutrinos, and the smaller one with solar neutrinos, as described in Section 4.2.

Figure 9.22 shows the zenith angle distribution of events attributed to muon and electron-neutrinos in the Superkamiokande detector, containing 50,000 tons of water viewed by 11,000 photomultipliers (see also Fig. 3.13). Charged current reactions of the electron- and muon-type neutrinos will result in production of charged electrons or muons. These emit Cerenkov radiation as they traverse the water (see Fig. 9.9), and this radiation appears as a ring of light at the water surface, which is detected by the photomultiplier array. Muons give clean Cerenkov rings, while those for electrons are more diffuse, because of bremsstrahlung and multiple scattering as the electron traverses the water (see also Fig. 1.1). The direction of the charged lepton is found from timing recorded by the photomultipliers, and at the energies involved, this gives a fair indication of the zenith angle of the incident neutrino.

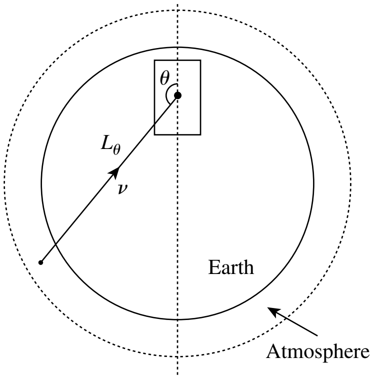


Fig. 9.23 Sketch illustrating the strong dependence of neutrino path lengths on the zenith angle θ. The dashed circle (not to scale!) indicates the typical height (20 km) of neutrino production in the atmosphere.

The typical path length of the neutrinos through the atmosphere and the Earth is a strong function of the zenith angle θ, as shown in Fig. 9.23. It ranges from about 20 km for neutrinos from directly overhead, to 200 km for those coming in horizontally, to 12,000 km for those coming vertically upwards from the atmosphere on the far side of the Earth. We might remark here that massive, multikiloton size detectors are necessary because of the smallness of the weak interaction cross-sections which are involved, and that on the other hand the very weakness of these interactions is being exploited in employing baselines going right through the Earth. The probability that a neutrino of the GeV energy range will be absorbed in such a diametral traversal is less than 0.1%.

The points in Fig. 9.22 show the observed rates of events above 1.3 GeV energy, the full-line histograms the event rates expected in the absence of oscillations, and the dashed histograms the best fits to an oscillation scenario. The electrons, and hence electron-neutrinos, show a zenith angle dependence consistent with no oscillations, while the upward-travelling muons from muon–neutrino interactions are strongly suppressed relative to the downward ones, the factor being ≈ 0.5 for those moving vertically upwards. In comparing these results to the expectations from Section 4.2, it is clear that because the events are integrated over a very broad energy spectrum as well as a range of path length, no actual oscillations will be observed and the mean value of the factor $\sin^2\left(1.27\Delta m^2 L/E\right)$ in (4.11) for large L will be just 0.5. Thus the fact that the observed suppression is 0.5 implies that $\sin^2(2\theta) \approx 1$, that is, the mixing is maximal as in Fig. 4.2 (see also Fig. 4.3). For the muon events, the best-fit histogram corresponds to a maximum mixing ($\theta_{23} = 45°$) and squared mass difference $\Delta m_{23}^2 \approx 2.3 \times 10^{-3} \left(\mathrm{eV}/c^2\right)^2$. Since the electron events show no zenith angle effect, these results are ascribed to $\nu_\mu \rightarrow \nu_\tau$ oscillations.

As pointed out in Chapter 4, the oscillation results first found for atmospheric neutrinos have later been confirmed in long baseline experiments at accelerators, notably the K2K experiment in Japan using a 250-km beam from the KEK laboratory to the Superkamiokande detector: the MINOS experiment with a 730-km (underground) beam from Fermilab (Chicago) to the Soudan mine in Minnesota; and the CNGS experiment using a 750 km beam from CERN to the Gran Sasso underground laboratory in Italy. These accelerator experiments employ detectors of higher resolution and much higher neutrino beam intensities (from the decay in flight of charged pions and kaons), and can measure the oscillation by comparing the spectrum of events in the far detector with that in a near detector placed close to the accelerator. They are of course now rapidly superseding the atmospheric neutrino detectors, which will in future be dedicated, for example, to study of geo-neutrinos (from radioactivity in the Earth), dark matter studies, and to Type II supernova neutrino watches (see Section 10.9).

9.16 Solar neutrinos

Anomalously low rates, which have also been interpreted in terms of neutrino oscillations, were first observed for solar neutrinos nearly 30 years ago, in pioneer experiments by Davis (1964, 1994) in the Homestake Mine, South Dakota, using a detector consisting of a tank filled with 615 tons of dry-cleaning fluid ($C_2\,Cl_4$), recording events due to the reaction

$$\nu_e + {}^{37}\mathrm{Cl} \rightarrow \mathrm{e}^- + {}^{37}\mathrm{A}$$

at the rate of about one argon atom per day. Section 10.3 describes the reactions involved in the so-called *pp* cycle of thermonuclear fusion of hydrogen to helium in the solar core. Neutrinos are produced from a number of reactions (see equations (10.6)–(10.11)).

Figure 9.24 shows the calculated fluxes of neutrinos at the Earth as a function of energy. Although the fluxes at the higher energies, notably from ^{8}B decay, are very small compared with those from the *pp* reaction, they make substantial contributions to the total event rate since the cross-sections for the detectors employed vary approximately as E_ν^3. Table 9.1 shows the results from several experiments, giving the ratio of the observed event rate to that calculated by Bahcall *et al.* (2001) in the absence of oscillations. The first two entries are for radiochemical experiments, detecting the accumulated activity of the product nuclei after fixed time periods. They offered formidable experimental challenges, requiring detection of less than one atom of the product element per day, in a mass of 50 tons (in the case of gallium) or 600 tons (in the case of chlorine in the Homestake experiment). The gallium experiments SAGE and GNO have a threshold energy of 0.2 MeV and are therefore sensitive to *pp* neutrinos, which from Fig. 9.24 extend up to 0.4 MeV energy.

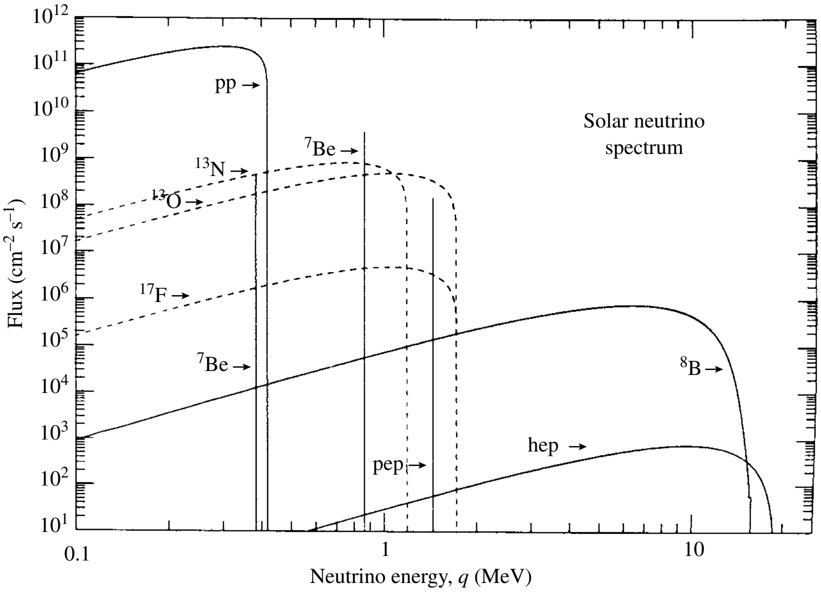


Fig. 9.24 Fluxes of solar neutrinos at the Earth from various reactions in the Sun (from Bahcall 1989).

Table 9.1 Solar neutrino experiments

Experiment		Reaction	Threshold (MeV)	Observed/ Expected Rate
SAGE + GNO	CC	^{71}Ga $(\nu_e, e)^{71}$Ge	0.2	0.58 ± 0.04
HOMESTAKE	CC	^{37}Cl $(\nu_e, e)^{37}$Ar	0.8	0.34 ± 0.03
SNO	CC	$\nu_e +^2 \mathrm{H} \rightarrow p + p + e$	~ 5	0.30 ± 0.05
SUPER-K	ES	$\nu + e \rightarrow \nu + e$	~ 5	0.46 ± 0.01
SNO	ES	$\nu + e \rightarrow \nu + e$	~ 5	0.47 ± 0.05
SNO	NC	$\nu + {}^2\mathrm{H} \rightarrow p + n + \nu$	~ 5	0.98 ± 0.09

CC = charged current (W-exchange); NC = neutral current (Z exchange); ES = electron scattering (via NC for ν_μ, ν_τ, and via NC and CC for ν_e)

The remaining experiments have higher thresholds and are not sensitive to the *pp* neutrinos. The SNO and SUPER-K experiments employ 1 kiloton of heavy water and 30 kilotons of light water respectively. Both measure events in real time, detecting the Cerenkov light emitted by the product electrons or γ-rays traversing the water using large photomultiplier arrays (see Figs. 4.7 and 9.25). The electrons originate from elastic scattering reactions (rows 4 and 5 of the table) or from charged current reactions (row 3). The SNO experiment also detects neutrons from disintegration of deuterium in a neutral current reaction (row 6 of the table). The neutron produced in this NC reaction is detected through capture by a deuteron in the heavy water, with emission of a 6.25 MeV γ-ray; or by adding 0.2% salt to the heavy water, resulting in neutron capture in chlorine with 8.6 MeV γ-rays; or by installing ^{3}He proportional counters to count neutrons directly. The typical threshold of 5 MeV for the SNO and SUPER-K experiments is determined by the radioactive background levels in the water, photomultipliers, and so on. The HOMESTAKE, SUPER-K, and SNO experiments are largely sensitive to the boron-8 neutrinos.

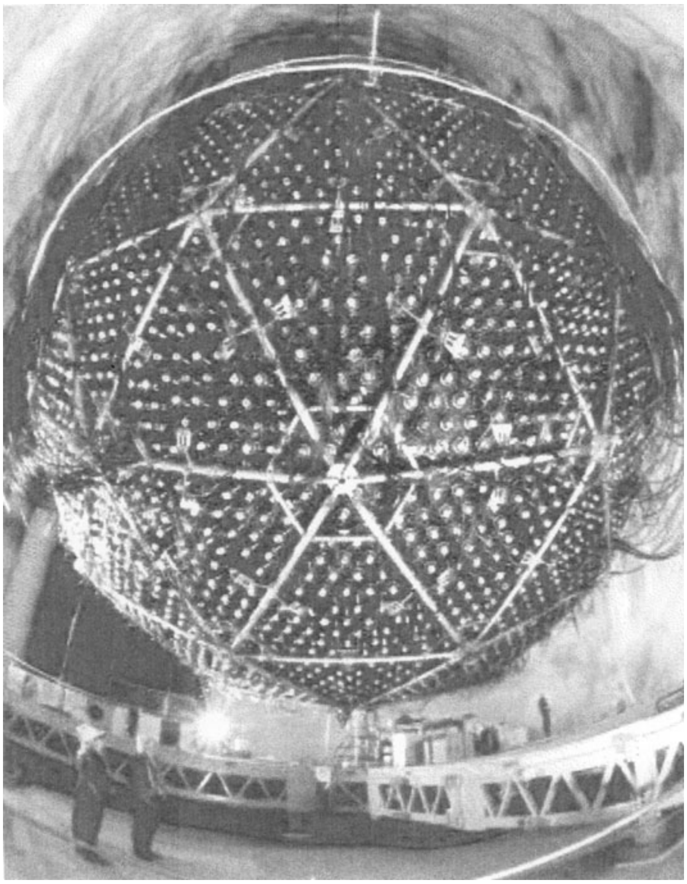


Fig. 9.25 The acrylic vessel employed to contain the 1 kiloton of heavy water in the SNO experiment, located in a mine at Sudbury, Ontario, together with some of the 9500 8″ photomultipliers used to record the Cerenkov light from electrons and γ-rays generated in interactions of solar neutrinos. The results on ν_e oscillations from solar experiments have been corroborated by long baseline experiments using $\bar{\nu}_e$ beams from reactors, for example the KAMLAND experiment in Japan.

The Superkamiokande experiment measures neutrino–electron elastic scattering, from the magnitude of the forward peak in the angular distribution of the scattered electrons relative to the Sun's direction (see Fig. 9.26). This can proceed through either charged current (CC) scattering (for ν_e only) *or* through neutral current (NC) scattering (via Z^0 exchange) which can apply to all flavours of neutrino, that is, to ν_e, ν_μ or ν_τ.

The significant facts about Table 9.1 are first, that the CC experiments measure a rate well below that expected, while the SNO neutral current reaction is consistent with expectations. Since the NC cross-section is the same for all neutrino flavours, the rate is independent of any oscillations the electron-neutrinos may undergo, and bears out the correctness of the solar model used to compute the fluxes in Fig. 9.24. Second, the ratio of observed to expected CC rates is less for the experiments insensitive to the low-energy (*pp*) neutrinos, implying an energy-dependent suppression factor. Finally, the SUPER-K and SNO results on electron scattering (ES) are in excellent agreement with the HOMESTAKE and SNO results on the CC reactions, if one accepts that these latter results indicate that only about 35% of the total neutrino flux arriving at the Earth is in the form of electron-neutrinos and that the remaining 65% has been transformed from ν_e to ν_μ and/or ν_τ, which then scatter from electrons through Z^0 exchange. The rate for these neutral current events calculated from the

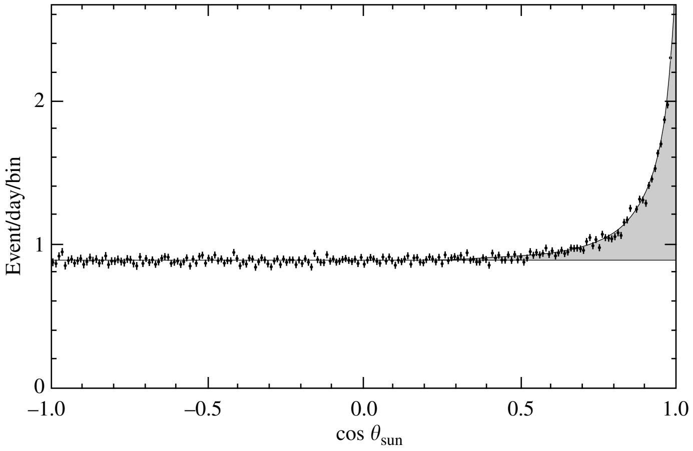


Fig. 9.26 Angular distribution of electrons relative to the Sun in the Superkamiokande experiment. The flat distribution in $\cos\theta$ is due to the background of secondary cosmic rays, while the forward peak is due to scattering by solar neutrinos, $\nu + e \rightarrow \nu + e$ (from Smy *et al.,* Superkamiokande collaboration 2002).

known value of the Weinberg angle ($\sin^2\theta_w = 0.23$) should then be about $1/3$ of that for the ν_e, bringing the expected total ES rate into excellent agreement with the measurement.

9.17 Neutrino oscillations in matter

The fact that the suppression factor for the CC reactions depends on the neutrino energy range, as evidenced by the first two entries in the above table, led to the possibility that mechanisms other than vacuum oscillations were involved. First Wolfenstein (1978) and later Mikhaev and Smirnov (1986) pointed out that the oscillations could be considerably modified by matter effects, namely by what is called the *MSW mechanism*, after the initials of these three physicists.

The MSW mechanism is described in Appendix D, and here we just outline the main points involved. Basically, an electron-neutrino ν_e produced via thermonuclear reactions (Section 10.3) in the solar core can, in its passage through the solar material, undergo both charged and neutral current interactions with electrons, while ν_μ and ν_τ neutrinos are limited to neutral current interactions only, since the energies involved, below 20 MeV, are insufficient to create the corresponding charged lepton. That means that ν_e are subject to an extra weak potential, equivalent to an increase in effective mass, and in turn this has the consequence that in a sort of resonant reaction, ν_e eigenstates can switch into ν_μ and/or ν_τ. These transitions depend on the solar electron density and the neutrino energy, and this has provided a correct explanation of why the suppression is greater for ^{8}B neutrinos (those above 1 MeV, see Fig. 9.24) than for the pp neutrinos, below 1 MeV. From the observed suppression factors as a function of neutrino energy, the relevant vacuum mixing angle θ_{12} can be calculated. It turns out to be quite large, as indicated in (4.12) and Fig. 4.3.

9.18 Point sources of high-energy neutrinos

The existence of point sources of TeV γ-rays described in Section 9.10 is usually attributed to electromagnetic processes, for example, synchrotron radiation by high-energy electrons accelerated in the source. However, it is also possible that γ-rays could be produced by pion production and decay, $\pi^0 \to 2\gamma$, and this would then imply high-energy neutrino point sources *via* $\pi^\pm \to \mu^\pm + \nu_\mu$. Such neutrinos could be detected in underground experiments *via* the secondary charged muons they produce. At TeV energies, the muon range in rock would be measured in kilometres, and the object would be to detect muons travelling *upwards* (since any downward flux would be completely swamped by atmospheric muons from pion decay in the atmosphere overhead). Because of the high energy, the direction of the secondary muon will follow rather closely that of the parent neutrino. Here the very weakness of their interactions is being exploited to detect neutrinos which have come up through the Earth and produced a signal free of background, except for the ubiquitous atmospheric neutrinos, which are, however, broadly distributed in angle.

The rate of such neutrino events (compared with that of γ-rays) is expected to be low on account of the weak cross-section. However, this is compensated, first by the fact that the cross-section for the interaction of a neutrino with a quark in the target nucleus rises as the square of the CMS energy, or linearly with the laboratory neutrino energy (see (1.27)); and that in the neutrino-quark collision, the secondary muon energy, and hence its range in rock, is also proportional to neutrino energy. So although the neutrino flux falls off rapidly with energy, it has to be multiplied by a factor of E_ν^2 to get the event rate. This argument holds up to TeV energies, but beyond that, the cross-sections flatten off because of the W propagator in (1.9), and the muon range is no longer proportional to energy because of radiation losses analogous to those of electrons discussed in Section 9.6, but coming in at a much higher energy (by a factor of order $(m_\mu/m_e)^2$).

The detection of rare high-energy cosmic neutrino interactions *via* the secondary upward-travelling muons has to be carried out on a large scale, and uses great depths of sea water or of ice, and again relies on the detection of the Cerenkov light radiated by the muon. Figure 9.27 shows one such experiment, in which strings of photomultipliers collect the Cerenkov signals over a volume with dimensions of order several hundred meters. They have of course recorded atmospheric neutrino events, but so far no evidence for high-energy neutrino point sources. These arrays may also detect high-energy neutrinos from possible WIMP–antiWIMP annihilations in the Sun (see Section 7.9).

9.19 Gravitational radiation

Of all the radiations incident upon the Earth from the cosmos, the most elusive and most difficult to detect is surely gravitational radiation. The key equations describing gravitational radiation, its production and detection are found from the general theory of relativity, which is outside the scope of this text. Nevertheless, we can understand many of the features of gravity waves by using the analogy with electromagnetic radiation.

First of all, we remark that in scattering problems it is usual to expand the function describing a plane wave into a superposition of spherical waves with different values of the orbital angular momentum, l, with respect to the scattering centre. An oscillation can equally be represented in the form of a multipole expansion, corresponding to a superposition of oscillators with

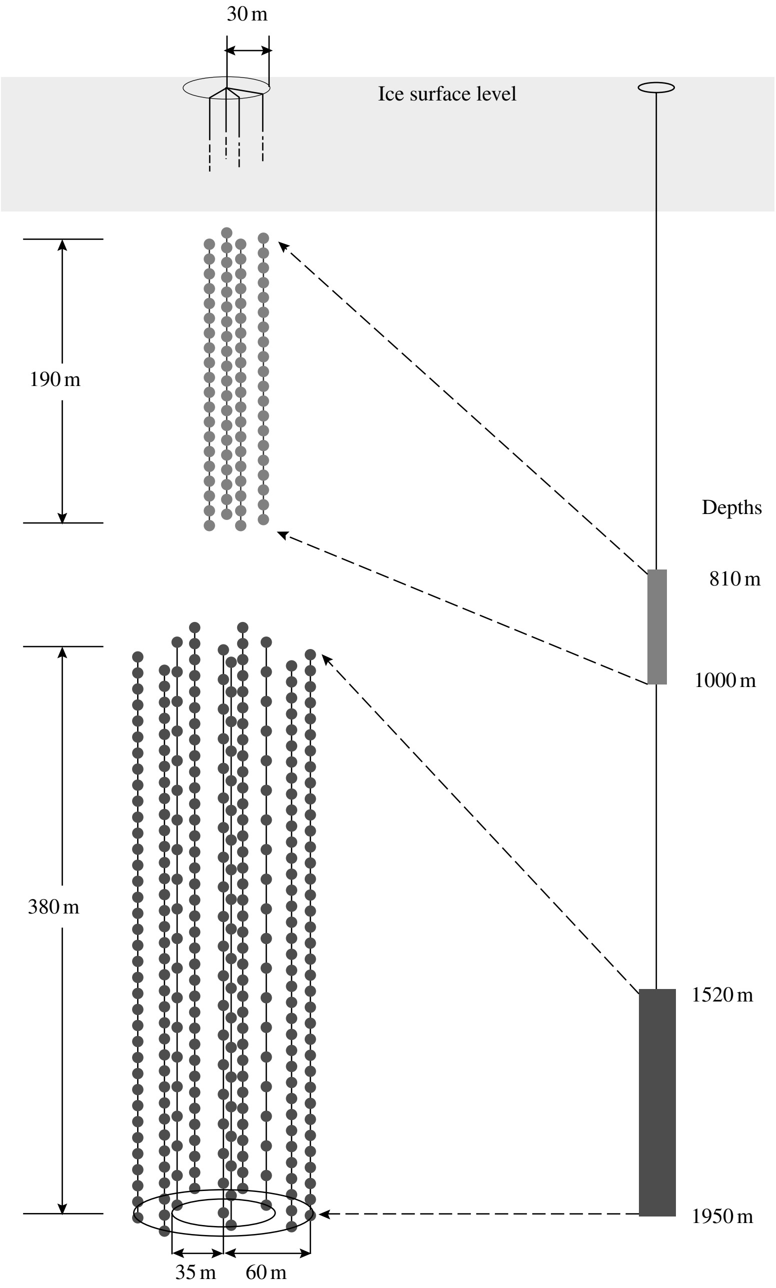


Fig. 9.27 Diagram of the strings of photomultipliers sunk in ice at the South Pole in the AMANDA experiment. They record Cerenkov light emitted by upward coming muons generated in neutrino interactions. Timing provides information on the zenith and azimuthal angles of the muon. The AMANDA array is currently being replaced by a much larger array, ICECUBE, covering 1 km^3 volume.

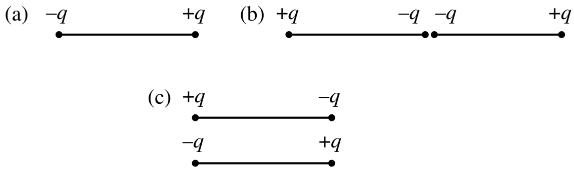


Fig. 9.28 (a) an electric dipole. (b) and (c) two configurations of an electric quadrupole, for which the dipole moment is zero.

different l-values. Thus dipole, quadrupole, sextupole, … oscillator terms are associated with $l = 1, 2, 3, \ldots$ waves. Figure 9.28 depicts a simple electric dipole and two versions of an electric quadrupole. If ω represents the angular frequency of the oscillation, the power radiated is given by the formula

$$P(l) = 2cF(l)\left(\frac{\omega}{c}\right)^{2l+2} |Q_{lm}|^2 \tag{9.24}$$

where $F(l) = (l+1)/\left[l\{(2l+1)!!\}^2\right]$. Here $n!! = 1.3.5.7\ldots$—and Q_{lm} is the l th moment of the distribution of electric charge density ρ, integrated over volume and projected along the z-axis, along which m is the component of angular momentum:

$$Q_{lm} = \int r^l\, Y^*_{lm}(\theta, \phi)\, \rho\, \mathrm{d}V \tag{9.25}$$

Here $Y_{lm}(\theta, \phi)$ is a spherical harmonic. In order of magnitude, an electric dipole moment ($l = 1$) has $Q_1 \sim er$ where r is the dimension of the system and e denotes the charge. Thus from (9.24) the power radiated is of order

$$P_{\text{dipole}} \sim \frac{\omega^4 e^2 r^2}{c^3} \tag{9.26}$$

The dependence on ω^4 arises because the power radiated is proportional to the square of the acceleration of the charges, varying as ω^2. The formula applies also to Rayleigh scattering of light by air molecules and by dust, and its dependence on $1/\lambda^4$ accounts for the blueness of the sky and the redness of the sunset. For an electric quadrupole moment ($l = 2$), $Q_2 \sim er^2$ and the power radiated is of order

$$P_{\text{quadrupole}} \sim \frac{\omega^6 e^2 r^4}{c^5} \tag{9.27}$$

The gravitational field is a tensor field (as compared with the vector, spin 1 photon field of electromagnetism), and gravitational interactions are mediated by gravitons of spin 2. As a result, dipole emission of gravitational waves is impossible and the simplest radiator of gravitational waves is an oscillating mass quadrupole. The power emitted can be estimated by replacing e^2 in the expression e^2/r for the electric potential between charges by $\mathrm{G}M^2$ in the expression $\mathrm{G}M^2/r$ for the gravitational potential between masses, so that

$$P_{\text{grav}} \sim \frac{\omega^6 \mathrm{G} M^2 r^4}{c^5} \tag{9.28}$$

Here, M denotes a typical mass and r a typical dimension in the quadrupole system, and we have just quoted the order of magnitude result for orientation. It should be emphasised that a quantitative calculation using Newtonian mechanics and simply substituting from the electric quadrupole formula will in any case underestimate the power in gravitational radiation, as computed using the general theory of relativity, by a factor 4. We quote here two examples of quantitative predictions. A rod of length L and mass M rotating about its mid-point with angular velocity ω emits gravitational radiation with the power

$$P = \left(\frac{2}{45}\right) \omega^6 \mathrm{G} M^2 \frac{L^4}{c^5} \tag{9.29}$$

while a binary system consisting of two stars in a circular orbit of diameter D and angular frequency ω radiates power

$$P = \left(\frac{32}{5}\right) \omega^6 \mathrm{G}\, \mu^2 \frac{D^4}{c^5} \tag{9.30}$$

where μ is the reduced mass, and for stars of equal mass M, $\mu = M/2$.

For an experiment on a laboratory scale, we could take $L \sim 1$ m, $M \sim 1$ kg and $\omega \sim 10^3\ \mathrm{s}^{-1}$, which from (9.28) gives $P \sim 10^{-36}$ watts only. According to (9.30), even the entire Earth, in its orbit around the Sun, radiates a mere 196 watts. More likely objects for a measurable level of gravitational radiation may be binary stars in a late stage of evolution, just as they coalesce to black holes, when a considerable fraction of the rest energy of the stars should be emitted as gravitational radiation. Indeed, it is believed that gravitational energy loss is what leads to the collapse. Somewhat less dramatic is the radiation emitted during the formation of a neutron star, which is followed by the spectacular optical display of a Type II supernova (see Sections 10.8–10.10).

9.20 The binary pulsar

The most convincing evidence for the existence of gravitational radiation, and through it, a quantitative test of general relativity, came from the observations of the binary pulsar PSR 1913 + 16 by Hulse and Taylor, first recorded in 1975. This binary consists of two neutron stars, of which one is a pulsar with a period of 0.059 s. Pulsars are neutron stars (see Section 10.10) which spin rapidly with frequencies of 10–100 s^{-1}, emitting radio waves in a beam which periodically sweep past the observer, like a rotating beam from a lighthouse. The orbital period of this binary, $\tau = 7.8$ h, and other characteristics of the orbit were found from the Doppler shift of the pulsar signal as this neutron star circulates its companion. The unique feature of this binary is that, during a period of more than 20 years of observation, a tiny but steady decrease in the orbital period was detected. After a few minor corrections for galactic accelerations the observed fractional rate of decrease in the period τ was found to be:

$$\frac{\mathrm{d}\tau}{\mathrm{d}t} = -\,(2.409 \pm 0.005) \times 10^{-12} \tag{9.31}$$

Such a decrease is to be expected if the pair lose energy by emission of gravitational radiation. To estimate the magnitude of this effect, let us assume for simplicity that the two stars have equal masses M and execute a circular orbit of diameter D. With $\omega = 2\pi/\tau$ as the angular frequency of the orbital motion, the tangential velocity of each star is $v = \omega D/2$. The total energy of the system is the sum of kinetic and potential energies:

$$E_{\text{tot}} = 2 \times \left(\frac{Mv^2}{2}\right) - \frac{\mathrm{G}M^2}{D} \tag{9.32}$$

Balancing the gravitational and centrifugal forces on one of the masses gives

$$2\frac{Mv^2}{D} = \frac{\mathrm{G}M^2}{D^2}$$

so that

$$E_{\text{tot}} = -\frac{\mathrm{G}M^2}{2D} \tag{9.33}$$

that is, the kinetic energy is just half the (absolute value of) the potential energy—another example of the virial theorem. Since from the above equations $E_{\text{tot}} \propto 1/D$ and $\omega \propto 1/D^{3/2}$, the orbital period $\tau \propto (E_{\text{tot}})^{-3/2}$. Hence

$$\left(\frac{1}{\tau}\right)\frac{\mathrm{d}\tau}{\mathrm{d}t} = -\left(\frac{3}{2E}\right)\frac{\mathrm{d}E}{\mathrm{d}t} = -\left(\frac{3}{2}\right)\frac{P}{E} \tag{9.34}$$

where P is the power radiated. The characteristics of the binary give $\omega = 2.2 \times 10^{-4}, M \sim 1.4 M_{\text{sun}}$. Inserting these values to obtain D and E_{tot}, and using (9.30) to estimate the power radiated, one obtains $\mathrm{d}\tau/\mathrm{d}t \sim 10^{-13}$. A full calculation, taking into account the eccentricity of the binary orbit and the inequality of the masses yields the result

$$\frac{\mathrm{d}\tau}{\mathrm{d}t} = -(2.4025 \pm 0.0002) \times 10^{-12} \tag{9.35}$$

agreeing exactly (within the 0.2% experimental error) with the observed value (9.31). This result has given great confidence that the basic physics is well understood and that, despite enormous experimental difficulties, the detection of gravitational radiation in the laboratory is worth pursuing.

Several such binary pulsars have now been observed. More remarkably, a double pulsar, PSR J0737-3039 A and B, that is, a binary system in which both neutron stars A and B are pulsars, has been discovered (Lyne *et al.* 2004). This system allows even more tests of general relativity, especially since fortunately, the plane of the orbit practically coincides with the line of sight. First, the collapse of the orbit due to gravitational radiation was observed, again in agreement with the predictions from general relativity. Second, an orbital precession (Section 2.8) of 17° per year—about 100,000 times larger than that for the orbit of the planet Mercury—was measured. The Shapiro delay (Section 2.7) in the signal was also detected as the beam from pulsar A passed close by pulsar B. Finally, the gravitational time dilation was observed, the pulse rate from one pulsar slowing down as it passed closer to the other. Again, we should emphasise that these tests of general relativity are wonderful but for relatively weak fields, far away from the non-linearities implied in the Einstein field equations for strong enough fields.

9.21 Detection of gravitational waves

When gravitational waves impinge on a detector, the difference in the acceleration from different parts of the wave can induce a deformation or strain, corresponding to an extension Δx in length x. The strain $h = \Delta x/x$ is given by

$$h^2 \sim \mathrm{G}\frac{P}{c^3\omega^2R^2} \tag{9.36}$$

where P is the power emitted by the source, R is its distance from the detector, and ω is the frequency of the radiation. Clearly, a detector with a quadrupole moment is necessary to excite a quadrupole amplitude. Inserting the value of P from (9.28) we find for the amplitude

$$h \sim \frac{\mathrm{G}ML^2\omega^2}{c^4R} \tag{9.37}$$

where the product $ML^2\omega^2$ is the second derivative of the quadrupole moment ML^2 of the source and is equal to the kinetic energy $E_{\mathrm{kin}} \sim Mv^2$ associated with the source oscillations. In a violent event such as the collapse to a neutron star, gravitational energy released due to the infall is of order $0.1\,M_{\mathrm{sun}}c^2$ (see Section 10.10). If we optimistically assume that 10% of this appears in the form of gravity waves then

$$h \sim \frac{GM_{\mathrm{sun}}}{100\,c^2R} \sim \frac{10^{-15}}{R} \tag{9.38}$$

where R is the distance of the source in parsec. For the local galaxy, $R \sim 10\,\mathrm{kpc}$ and $h \sim 10^{-19}$, while for the Virgo cluster of galaxies $R \sim 10\,\mathrm{Mpc}$ and $h \sim 10^{-22}$. Note that, even for a bar 1 km long, $h = 10^{-19}$ corresponds to a change in length of 10^{-16} m or one-tenth of a nuclear radius! It is likely that, by going further afield, the decrease in h with increasing R in (9.38) may be compensated by the R^3 increase in the number of sources and the possibility of much more violent events, such as collapse to massive black holes (AGNs), with gravitational wave energy far exceeding the solar mass energy.

Although the conceivable distortions to be measured by gravitational wave detectors as a result of the most violent cosmic events will be, at best, of order 10^{-20}, they are not considered to be beyond reach. The technique for their detection is based on split laser beams and a Michelson interferometer—see Fig. 9.29. The laser light is split into two paths at right angles by the beam splitter B. The beams are reflected back and forth by mirrors M1 – M4 (M1 and M3 being half-silvered) attached to masses and the fringes observed when the light beams recombine and interfere. A gravitational wave will effectively stretch one dimension, say $D1$, and contract the orthogonal dimension $D2$, thus causing a

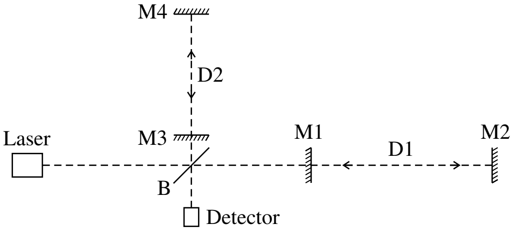


Fig. 9.29 Sketch of a Michelson interferometer layout for a gravitational wave detector.

fringe shift; alternatively, one can think of the gravitational wave as distorting the space/time between the mirrors, thus introducing tiny (Shapiro) delays in the relative transit times of the light beams. A Fabry-Perot etalon is used so that the beams make many traverses to and fro before recombining. As an example, the values chosen for the two LIGO experiments in the USA are $D1, D2 \sim 4\,\mathrm{km}$, operating in a frequency range 10–1000 Hz, typical of the collapse times for neutron stars/black holes. In Europe, the dimensions of the VIRGO experiment are similar, while GEO600 is a smaller array with $L = 0.6\,\mathrm{km}$. All of these experiments have very comparable sensitivities, and all are extremely difficult, and require years of tuning before scientific runs can commence.

The main problems for these experiments are the effects of background (seismic) noise, which kills all hope of detecting a signal below about 10 Hz. The noise problem can be tackled to some extent by combining the signals from two or more of the several detectors located in different positions worldwide, using timing information to reduce the noise and also indicate the direction of the source. For example, one of the LIGO and the TAMA (Japan) detectors have operated in coincidence for almost 500 h and placed an upper limit on any signal of < 0.12 events per day (Abbott *et al.* 2006). The feasibility of combining signals from all five of the above detectors has been proven, and serious scientific runs are commencing. More ambitious plans, aimed at avoiding low frequency (seismic) noise, are to place several gravity wave detectors in Earth orbit (codenamed LISA). There is no question that, with incremental improvements in the detection systems, gravitational waves will eventually be detected.

9.22 Summary

- The charged primary cosmic rays consist principally of high-energy nuclei of the elements, their chemical composition being in general similar to the solar system abundances. The exception is for lithium, beryllium, and boron, which are abundant in the cosmic rays and produced by spallation of heavier nuclei in collisions with interstellar matter.
- The energy spectrum up to 10^{15} eV falls off as a power law, $\mathrm{d}N/\mathrm{d}E \sim E^{-2.7}$, and decreases more rapidly at higher energy, up to at least 10^{20} eV.
- The charged primary radiation is affected by Solar System magnetic fields. The Earth's field imposes a cut-off in momentum depending on magnetic latitude. Cosmic rays are also moderated by solar effects (the solar wind) which follows the 11-year sunspot cycle.
- The energy density in cosmic rays, at about $1\,\mathrm{eV\,cm^{-3}}$, is comparable with that in the cosmic microwave background, in starlight and in galactic magnetic fields. The rate at which energy needs to be injected into the cosmic rays can be accounted for in terms of shock-wave acceleration in supernova shells, provided these processes have efficiencies of a few percent. While this mechanism can work up to energies of 10^{14} eV, the acceleration mechanism for the highest energies is unknown.
- The cosmic rays at sea-level are of secondary origin, and generated by collisions of the primaries in the atmosphere. The hard component consists of muons from decay of charged pions created in the atmosphere,

while the soft component consists of electrons and photons originating from the decay of neutral pions.

- High-energy cosmic ray nuclei can generate a nuclear cascade in the atmosphere, and this can lead to an extensive air shower, consisting of nucleons, muons, and electron–photon cascades extending over a large area (typically of radius 1 km). Practically all of the energy at sea-level is in the electron–photon component, and there is a linear relation between the primary energy and the shower size.
- The electron–photon showers can be detected via the Cerenkov light or scintillation light they generate in traversing the atmosphere.
- At energies above 10^{19} eV, interactions of the primaries with the microwave radiation, leading to pion production, is expected to show suppression effects (the GZK cut-off), and both the HiRes and AUGER experiments have observed this.
- Point sources of γ-rays of energies up to 30 GeV have been detected with the EGRET detector on the GRO satellite. The sources include pulsars and AGNs. Point sources involving γ-ray energies in the TeV region and above have been detected using the ground-based air Cerenkov method.
- Intermittent as well as steady sources are detected. The sporadic sources consist of bursts lasting 10 ms to 10 s, which can disappear completely and reappear a year or so later. In the TeV energy region the γ-ray bursts are known as blazars. The shortness of the bursts indicates compact sources, and the blazar rate is consistent with the estimated rate of mergers of binary neutron stars to form black holes.
- Atmospheric neutrinos from decay of pions produced in the atmosphere have been studied extensively in deep underground experiments, and the studies show clear evidence for oscillations in the neutrino flavour (ν_μ or ν_e) over baselines comparable with the Earth's radius, associated with the differences of mass of the neutrino mass eigenstates. The amplitude of the ($\nu_\mu \to \nu_\tau$) mixing is near maximal.
- Similar oscillation phenomena have been observed for neutrinos from the Sun. The suppression of ν_e events due to mixing shows an energy dependence, which can be described in terms of matter-induced oscillations inside the Sun.
- Attempts are under way to detect point sources of high-energy neutrinos.
- The existence of gravitational radiation of the expected magnitude has been demonstrated from the slow-down rate of binary pulsars. Attempts to detect gravitational radiation directly are currently under way.

Problems

More challenging problems are marked with an asterisk.

(9.1) Relativistic cosmic ray protons are accelerated by a shock front. Deduce the form of the differential energy spectrum of the protons, assuming that the probability that a proton will re-cross the front is 80% and that the fractional increase in energy per crossing is 20%.

(9.2) The refractive index, n, of air at sea-level is given by $n - 1 = 2.7 \times 10^{-4}$, a quantity which is proportional to pressure. Calculate the radial spread in metres of the ring of Cerenkov light at sea-level, due to an ultra-relativistic charged particle travelling vertically downwards at a depth of 100 gm cm^{-2} in the atmosphere. Assume an exponential atmosphere with density ρ and height h related by $\rho = \rho_0 \exp(-h/H)$ where $H = 6.5$ km. The total atmospheric depth is 1030 gm cm^{-2}.

(9.3) The average rate of energy loss of ultra-relativistic muons of energy E in traversing x gm cm^{-2} of material is given by the formula $\mathrm{d}E/\mathrm{d}x = a + bE$, where $a = 2.5$ MeV gm^{-1} cm^{2} is the rate of ionization loss and the second term accounts for radiation energy loss. Calculate the average range in kilometres of a 5000 GeV muon in rock of density 3 gm cm^{-3}, for which the critical muon energy is 1000 GeV.

*(9.4) Primary cosmic ray protons interact in the atmosphere with mean free path $\lambda \sim 100$ gm cm^{-2}. They produce relativistic charged pions of energy E travelling vertically downwards. These pions may subsequently decay in flight, or they may undergo nuclear interaction, again with a mean free path equal to λ. Assuming an exponential atmosphere of scale length H, show that the overall probability that a pion will decay rather than interact is $P = E_0/(E_0 + E)$ where $E_0 = m_\pi c^2 H/c\tau_\pi$. Calculate the value of E_0. How is the expression for P modified if the pion is produced at angle θ to the zenith? The total depth of the atmosphere (1030 gm cm^{-2}) can be assumed to be very large compared with λ($m_\pi c^2 = 0.14$ GeV: $H = 6.5$ km: $\tau_\pi = 26$ ns).

(9.5) State whether you believe that CP violating effects in a neutrino beam are possible if the mixing is between just two flavour eigenstates. What happens if matter effects are taken into account, for a beam traversing the Earth?

*(9.6) In the Kamiokande experiment, solar neutrinos are observed through the process of elastic scattering of electrons, and the detection of the Cerenkov light emitted by the recoiling electron as it traverse the water detector. If the incident neutrino has energy E_0, calculate the angle of scattering of the electron in terms of its recoil energy E. (Assume that the electron (and neutrino) masses can be neglected in comparison with the energies.)

*(9.7) High-energy charged pions decay in flight in the atmosphere. Calculate the mean fractional energy received by the muon and by the neutrino in the decay $\pi^+ \to \mu^+ + \nu_\mu$. Estimate also the mean fractional energy of the pion, which is carried by each of the neutrinos (antineutrinos) in the subsequent muon decay $\mu^+ \to e^+ + \nu_e + \bar{\nu}_\mu$. Assume that all neutrinos are massless and neglect ionization energy losses in the atmosphere and polarization effects in muon decay ($m_\pi^2 c = 0.139$ GeV, $m_\mu c^2 = 0.106$ GeV).

(9.8) Calculate the power radiated by the Moon in the Earth's orbit due to gravitational radiation. (the Earth's mass $= 6 \times 10^{24}$ kg, the Moon's mass $= 7.4 \times 10^{22}$ kg, the mean Earth–Moon distance $= 3.8 \times 10^5$ km, orbital period of the Moon $= 27.3$ days).

(9.9) Use the formulae and data in Section 9.20 to check the estimate of the decrease in orbital period of PSR 1913 + 16, assuming a circular orbit and equal masses for the members of the binary star system.

(9.10) Estimate the fraction of muon–neutrinos of energy 1 GeV which interact in traversing the Earth's diameter (take theEarth's mean density as 3.5 gm cm^{-3}, and the Earth 's radius as 6400 km).

(9.11) Show that, when a relativistic charged particle, travelling in the positive x-direction, is scattered backwards by the field due to a shock front moving with non-relativistic velocity u_1 in the negative x-direction, it receives a fractional energy increase of order u_1/c.

(9.12) Show that neutrinos, originating in the atmosphere at altitude h, and at zenith angle θ lying between 0 and π, have path length L to a detector placed near the Earth's surface given by $L = (R^2 \cos^2\theta + R^2 + Rh)^{1/2} - R\cos\theta$. Hence verify the values of L quoted in Section 9.15.

*(9.13) Calculate the apparent transverse velocity of a jet travelling with velocity v at inclination θ to the line of sight to the Earth, and show that this transverse velocity can appear to be superluminal, with a maximum value $\gamma\beta$c, where $\beta = v/c$ and $\gamma = 1/\left(1 - \beta^2\right)^{1/2}$.

Particle physics in stars and galaxies

10

10.1 Preamble

In previous chapters we have traced the primordial development of the universe, principally through the properties and interactions of the elementary particle constituents of matter. This phase came to an end when the temperature of the fireball fell below a value $kT \sim 0.3\,\text{eV}$ and $z \sim 10^3$, when the radiation and (baryonic) matter decoupled, and not long after the time when the universe became matter dominated. The previously opaque universe, consisting of a plasma of electrons, hydrogen and helium nuclei, and photons, was then replaced by a relatively transparent but almost totally dark universe, consisting very largely of clouds of neutral atoms and molecules. From these clouds, stars were able to form by gravitational infall as soon as the redshift fell to $z \sim 12$. Then there was light. The process of star formation has of course been continuous since that time.

The evolution of the stars, at least during most of their life, is only peripherally linked with elementary particle physics as such, so we shall discuss it rather briefly and in particular contexts, for example, with respect to solar neutrinos described in Chapter 9. However, the later stages of stellar evolution do depend very directly on particle processes at a fundamental level and at quite high energies, involving some of the most violent events in the universe, and we describe these in more detail.

10.2 Stellar evolution—the early stages

As described in Chapter 8, stars can condense out of gas clouds (predominantly hydrogen) once the mass and density of the material fulfils the Jeans criterion (8.37). The gravitational potential energy lost under contraction goes to heat up the gas. The resulting gas pressure opposes further contraction, and the so-called protostar reaches a state approaching hydrostatic equilibrium (see Example 10.2 below). Typically, the gas density at this stage is $10^{-15}\,\text{kg m}^{-3}$ and the radius is 10^{15} m (i.e. about one million times the solar radius). As the star radiates energy from the envelope it slowly contracts further, to about 100 times the solar radius. By then, the amount of gravitational energy that has been released is of order 10 eV per hydrogen atom, so that collisional dissociation of hydrogen molecules (requiring 4.5 eV) and ionization of hydrogen atoms (13.6 eV) can take place. An equilibrium between photons, ionized and unionized matter arises, and the energy generated when the atoms and molecules recombine is released as photons to the outside world, allowing still further contraction. Without any

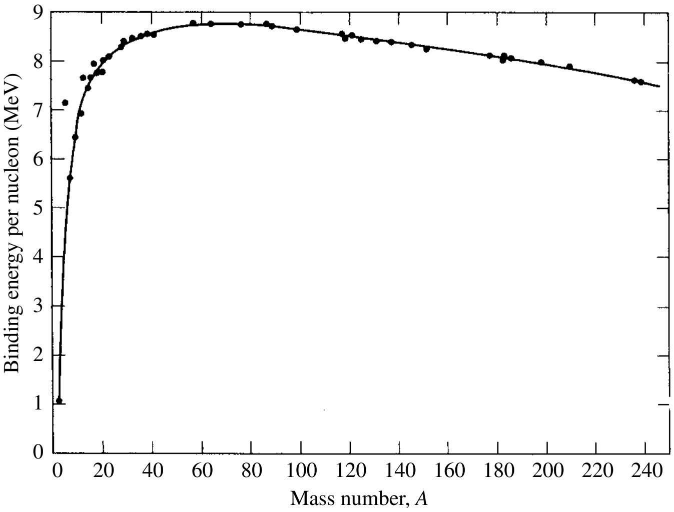


Fig. 10.1 Binding energy per nucleon as a function of mass number A, for nuclei stable against beta decay. The maximum binding is in the Fe–Ni region of the Periodic Table (Enge 1972).

source of energy apart from gravitation, the energy radiated by the star must always be compensated by further contraction, and a consequent increase in pressure and temperature of the core. In fact the kinetic (heat) energy of the star must be just equal to half its gravitational energy, an example of the *virial theorem* on the partition of kinetic and potential energy in a non-relativistic system of particles in thermal equilibrium, bound by an inverse square law potential (see Section 7.2).

Further collapse of the star is eventually halted by the onset of *thermonuclear reactions.* Figure 10.1 shows the binding energy per nucleon as a function of the mass number A of the nucleus. It is clear that if two light nuclei fuse to form a heavier nucleus, energy will be released, provided the product nucleus has $A < 56$, the mass number of iron, for which the binding energy per nucleon is a maximum. The amount of energy released is substantial. For example, if as described below, helium is formed from hydrogen, the binding energy liberated is of order 7 MeV per nucleon.

The electrostatic potential between two nuclei having charges Z_1e and Z_2e and mass numbers A_1 and A_2 with separation r is $V = Z_1Z_2e^2/4\pi r$. When just in contact, $r = r_0\,(A_1^{1/3} + A_2^{1/3})$ where $r_0 = 1.2\,\text{fm}$ is the unit nuclear radius. The first stage of the p–p fusion process in the Sun is the weak reaction

$$p + p \to d + \mathrm{e}^+ + \nu_e + 0.32\,\text{MeV} \tag{10.1}$$

with a Coulomb barrier height $V_0 = (1/4\pi)\left(e^2/2r_0\right) = \left(e^2/4\pi\hbar c\right) \times (\hbar c/2r_0)$. With $e^2/4\pi\hbar c = \alpha = 1/137$, $\hbar c = 197$ MeV fm and $r_0 = 1.2\,\text{fm}$ one finds $V_0 = 0.6\,\text{MeV}$. This is very much greater than the thermal energy of protons at the core temperature of the Sun, which can be estimated from the solar luminosity to be $kT \sim 1\,\text{keV}$. Although in classical terms the two nuclei cannot therefore surmount the Coulomb barrier, in quantum mechanics they can penetrate *through it* with finite probability. This effect had, in the 1920s, successfully accounted for the long lives of radioactive nuclei undergoing alpha decay. The barrier penetration probability is given by the approximate formula, only valid for $E \ll E_G$:

$$P(E) = \exp\left[-\left(\frac{E_G}{E}\right)^{1/2}\right] \tag{10.2}$$

where

$$E_G = \left(\frac{2m}{\hbar^2}\right)\left(\frac{Z_1 Z_2 e^2}{4}\right)^2 \tag{10.3}$$

is the so-called Gamow energy (named after George Gamow, who first investigated the barrier penetration problem). The quantity m is the reduced mass of the two nuclei, and for the *p–p* reaction it is half the proton mass, $m_p/2$. With $e^2 = 4\pi\alpha\hbar c$ one gets $E_G = m_p c^2 \pi^2 \alpha^2 = 0.49\,\text{MeV}$. So, if the relative kinetic energy $E \sim 1\,\text{keV}$ as indicated above, the barrier penetrability will be of order $P \sim \exp(-22) \sim 10^{-10}$. In fact the protons will have a Maxwellian distribution in kinetic energy of the form

$$F(E)\,\mathrm{d}E \sim E^{3/2} \exp\left(-\frac{E}{kT}\right)\mathrm{d}E \tag{10.4}$$

As shown in Fig. 10.2, the penetrability factor in (10.2) increases with energy, while for $E > kT$, the number of protons in (10.4) decreases with energy. The fusion rate is the product of these two distributions. However, even if the barrier is successfully penetrated, the usual reaction between the protons will be elastic scattering (via the strong interaction) rather than the weak reaction (10.1), which turns out to have a relative probability of order 1 in 10^{20}. So, although any one proton in the Sun is having millions of encounters with other protons every second, the average time for the conversion of protons to deuterium and helium nuclei is billions of years. The rate of nuclear energy generation is exactly matched to the energy radiated from the solar envelope, and while the hydrogen fuel lasts, the Sun is quite stable against any fluctuations in radius. Thus if the radius were to increase slightly, the surface area and energy radiated would increase, and this must be matched by an increase in fusion energy and core temperature, which in turn leads to contraction of the envelope. It is of interest to remark here that the nuclear energy generated in the core takes a very long time to work its way out to the photosphere (see Example 10.1).

Example 10.1 *Estimate the time required for energy generated by fusion in the solar core to reach the photosphere by radiative diffusion, given that the core temperature is 16 million °K, the surface temperature is 6000 °K and the solar radius is $R = 6 \times 10^8$ m.*

At the core temperature, the energy per photon is $\sim$ 1 keV, thus in the X-ray region. The radiation is transmitted to the surface through random collisions with the plasma particles, resulting in scattering, absorption, and re-emission processes. If we denote the steps between collisions by $\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3, \ldots$ the mean square distance travelled in this 'random walk' will be $L^2\,(\text{av}) = \left\langle (d_1 + d_2 + d_3 + \cdots + d_N)^2 \right\rangle = Nd^2$ where d is an average step length and N is the number of steps, and we have used the fact that the steps are in random directions so that in evaluating the square, all the cross-terms cancel. Thus to reach the surface from the deep interior

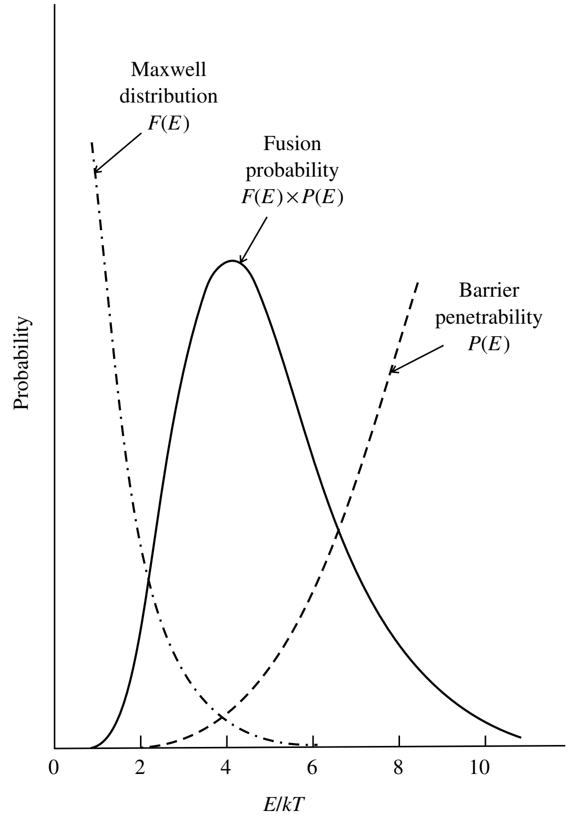


Fig. 10.2 Curves showing at left the Maxwell distribution of relative energy of colliding nuclei, and at right the barrier penetrability, for the *p–p* reaction. The fusion rate is proportional to the product of these distributions and is shown by the solid curve.

requires $L \sim R$ and $N \sim R^2/d^2$ steps, with an elapsed time $t_1 \sim R^2/cd$. Had the radiation been free to escape directly, the time to the surface would only have been $t_2 = R/c$, so that the process of radiative diffusion has slowed down the rate at which energy escapes the Sun by a factor $t_1/t_2 = R/d$. This is the factor by which the core luminosity, of order $R^2T_c^4$, is reduced to the surface luminosity, of order $R^2T_s^4$. Thus $d/R \sim (T_\mathrm{s}/T_\mathrm{c})^4$ and $t_1 = (R/d)\,(R/c) \sim 10^{14}$ s or about a million years.

10.3 Hydrogen burning: the p–p cycle in the Sun

The production of energy in the Sun is via the fusion of hydrogen to helium, according to the net process

$$4p \to {}^4\mathrm{He} + 2\mathrm{e}^+ + 2\nu_e + 26.73\ \mathrm{MeV} \qquad (10.5)$$

This process takes place in several stages. The first is the weak reaction (10.1) forming a deuteron:

$$p + p \to d + \mathrm{e}^+ + \nu_e \qquad (10.6)$$

which also has a small (0.4%) contribution from the so-called ‘pep’ process

$$p + e^- + p \to d + \nu_e$$

The next stage is the electromagnetic reaction

$$p + d \to {}^3\mathrm{He} + \gamma \tag{10.7}$$

followed by two strong interaction processes. The first produces ^{4}He according to

$$^3\mathrm{He} + {}^3\mathrm{He} \to {}^4\mathrm{He} + 2p \qquad (85\%) \tag{10.8}$$

while the second produces heavier elements, generating ^{4}He in subsequent interactions:

$$^3\mathrm{He} + {}^4\mathrm{He} \to {}^7\mathrm{Be} + \gamma \tag{10.9}$$

$$\mathrm{e}^- + {}^7\mathrm{Be} \to {}^7\mathrm{Li} + \nu_e; \quad p + {}^7\mathrm{Li} \to 2\,{}^4\mathrm{He} \quad (15\%\,) \tag{10.10}$$

$$p + {}^7\mathrm{Be} \to {}^8\mathrm{B} + \gamma; \quad {}^8\mathrm{B} \to {}^8\mathrm{Be}^* + \mathrm{e}^+ + \nu_e;$$

$$^8\mathrm{Be}^* \to 2\,{}^4\mathrm{He} \quad (0.02\%) \tag{10.11}$$

where the percentages indicate the contributions to the total helium production rate.

In addition to the *p–p* cycle, another cycle involving the elements C, N, and O accounts for about 1.6% of helium production in the Sun. Since the nuclear charges are larger, this cycle is more important for more massive hydrogen burning stars with higher core temperatures, such as Sirius A.

The above reactions have been written out in detail, because the study of neutrinos from these reactions has been crucially important, not only in verifying our picture of solar fusion reactions but also in demonstrating the existence of neutrino flavour oscillations, as has been described in Section 9.16.

Assuming that protons in the core account for about 10% of all protons in the Sun, the observed solar luminosity (3.8×10^{26} J s^{-1}) implies through equation (10.5) that an average survival time before a proton undergoes fusion is several billion years, comparable indeed with the age of the universe. This time is determined by several factors: the Coulomb barrier penetration rate; the weak interaction cross-section in the first stage of the *p–p* cycle; and most importantly, the opacity of the solar material, which determines the rate at which energy can escape from the Sun, and hence the rate at which fusion energy is generated in the core.

As indicated above, a star first lights up when it contracts from a ‘protostar’ and commences hydrogen burning, which it continues for most of its life. In the Herzsprung–Russell diagram of luminosity versus surface temperature (see Fig. 10.3) a hydrogen burning star is situated at a single point on a curve called the *main sequence*. The position on the curve depends on the mass, which determines the luminosity. Empirically, the luminosity L of main sequence stars varies with mass M as $L \propto M^{3.5}$, and so the lifetime τ of a star on the main sequence must be proportional to M/L, so $\tau \propto M^{-2.5}$. Those stars in the cluster above a certain mass, corresponding to $\tau = t_0$, the age of the cluster, have already moved off the main sequence at the main sequence turn-off (MSTO), towards the *red giant branch*, as further contraction of the helium core is halted by electron degeneracy pressure (see Section 10.6). Hydrogen burning still proceeds in a shell outside of the helium core, the envelope expands and the surface temperature falls. Eventually, as the core temperature rises, the helium core ignites and the star moves over to the *horizontal branch*. If the carbon core in turn ignites (i.e. if the star is massive enough), the star moves back again towards the red giant branch. If M is only of order one solar mass, however, no fusion of the carbon/oxygen core takes place, and the star moves down towards the left-hand bottom of the diagram where white dwarfs (not shown) are situated.

The age of the cluster—and hence the age t_0 of the universe for the very oldest clusters—can be found from the position of the MSTO and a stellar evolutionary model. In this way the age of the universe has been estimated at $14 \pm 2\,\text{Gyr}$.

10.4 Helium burning and the production of carbon and oxygen

When most of the hydrogen in the stellar core has been converted to helium, and fusion energy is no longer produced there, the core contracts and part of the gravitational energy released leads to local heating, the rest escaping from the core. The higher central temperature means that hydrogen burning now proceeds in a spherical shell surrounding the helium core, so the total mass and density of helium increases. If the star is sufficiently massive (more than about half a solar mass), the core temperature becomes high enough to ignite the helium, at a temperature just above 10^8 K, resulting in production of carbon and oxygen, the most abundant elements in the universe after hydrogen and helium.

Helium burning involves the following somewhat complex chain of events. In the absence of any stable nuclei with masses 5 or 8, fusion has to proceed by the so-called *triple alpha* reaction, as first discussed by Salpeter in 1952. The first stage is

$$^4\text{He} + {}^4\text{He} \longleftrightarrow {}^8\text{Be} \tag{10.12}$$

The nucleus ^{8}Be is unstable by 92 keV, so that the two helium nuclei must have this relative energy in order to 'hit' the ground state, and to do this efficiently requires a temperature, read off from a curve of the same type as shown in Fig. 10.2, of $T = (1-2) \times 10^8$ K. The mean lifetime of ^{8}Be is about 2.6×10^{-16} s. The equilibrium concentration of beryllium nuclei in (10.12) is only about one billionth of that of the helium nuclei. Nevertheless, a ^{8}Be nucleus may capture a third alpha particle to form the nucleus $^{12}\text{C}^*$ in an excited state at 7.654 MeV, which is just 0.3 MeV above the threshold energy for $^4\text{He} +^8$ Be:

$$^4\text{He} + {}^8\text{Be} \longleftrightarrow {}^{12}\text{C}^* \tag{10.13}$$

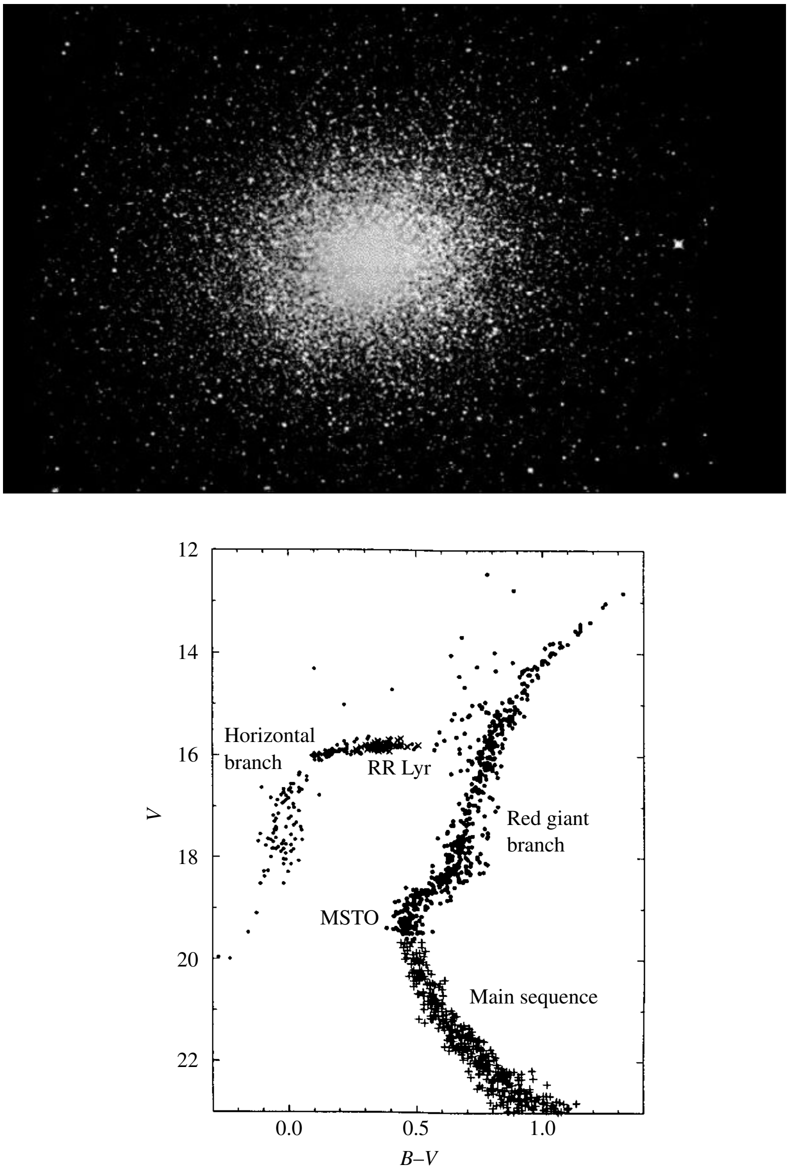


Fig. 10.3 (a) The globular cluster M13, containing upwards of 10^5 stars. Our galaxy contains at least 150 such globular clusters, in orbit about the centre of the galaxy. The stars in a globular cluster were formed together, and they orbit about the cluster centre, to which they are gravitationally bound. Such clusters are extremely old, and as described below, can be used to set an estimate of the age of the universe. (b) The Herzsprung–Russell diagram of stellar luminosity against surface temperature (colour), for stars in the globular cluster M15. Stars in such a cluster were all formed at essentially the same time, at a very early stage of the universe. In this graph, the magnitude (or logarithm of the luminosity) observed with a V filter (λ = 540 nm) is plotted vertically against the difference in magnitudes with a B filter (λ = 440 nm) and a V filter. Redder colours and cooler surface temperatures are to the right, and higher luminosities or lower magnitudes towards the top (from Chaboyer 1996).

Hoyle (1954) had pointed out the need for, and estimated the energy of, this resonant state in carbon, and it was subsequently found in accelerator experiments. Usually, the carbon decays back to beryllium plus helium as in (10.13), but can, with only 3×10^{-4} probability, decay by radiative transitions to the ground state:

$$^{12}\text{C}^* \longrightarrow {}^{12}\text{C} + 2\gamma \quad \left(\text{or e}^+ + \text{e}^-\right) \tag{10.14}$$

Once carbon nuclei have been created, the next stage of oxygen production can proceed via radiative alpha particle capture:

$$^{4}\text{He} + {}^{12}\text{C} \longrightarrow {}^{16}\text{O} + \gamma \tag{10.15}$$

It is fortunate that in this case there is *no* resonance in ^{16}O anywhere near the threshold energy, so that not all the carbon is consumed as soon as it is produced, and both carbon and oxygen are abundant elements in the universe. Obviously, the existence of the 7.654 MeV resonance level in the carbon-12 nucleus was vital for the development of carbon-based biological molecules and life as we know it in our particular universe.

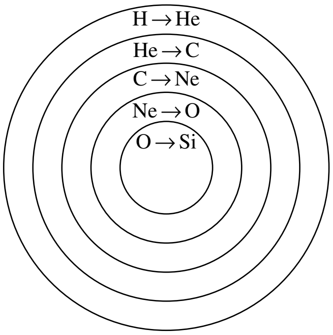


Fig. 10.4 Onion-like appearance of cross-section of massive star at an advanced stage of nuclear fusion. The heaviest elements are in the core, where the temperature and density are greatest, and are surrounded by lighter elements in spherical layers of successively lower temperatures and density.

10.5 Production of heavy elements

A massive star evolves further through fusion reactions to produce successively heavier elements, involving higher Coulomb barriers and higher core temperatures. A cross-section through the star would have an onion-like appearance, as in Fig. 10.4, with the heaviest elements in the core and lighter ones in spherical shells of successively larger radius and lower temperature.

Carbon burning commences when the core temperature and density are $T \sim 5 \times 10^8$ K and 3×10^9 kg m^{-3} respectively, and leads to the production of neon, sodium, and magnesium nuclei:

$$\begin{aligned} {}^{12}\mathrm{C} + {}^{12}\mathrm{C} &\longrightarrow {}^{20}\mathrm{Ne} + {}^{4}\mathrm{He} \\ &\longrightarrow {}^{23}\mathrm{Na} + p \\ &\longrightarrow {}^{23}\mathrm{Mg} + n \end{aligned} \tag{10.16}$$

At still higher temperatures, of order 2×10^9 K, oxygen burning leads to the production of silicon:

$${}^{16}\mathrm{O} + {}^{16}\mathrm{O} \longrightarrow {}^{28}\mathrm{Si} + {}^{4}\mathrm{He} \tag{10.17}$$

At such temperatures, the thermal photons have appreciable quantum energies. For example, a tiny proportion ($\sim 10^{-18}$) of the photons, with over 20 times the mean energy, will have energies above 9 MeV and can therefore cause photodisentegration of silicon, with the important production of helium nuclei:

$$\gamma + {}^{28}\mathrm{Si} \longrightarrow {}^{24}\mathrm{Mg} + {}^{4}\mathrm{He} \tag{10.18}$$

On account of their lower Coulomb barriers, the helium nuclei released can now, by radiative capture, induce successive fusions to form sulphur, argon, calcium, and eventually iron and nickel. These reactions proceed easily and the overall rate is really determined by the first stage of photoproduction (10.18). With the production of ^{56}Fe however, the exothermic fusion process finally ends, since as indicated in Fig. 10.1, iron is the most strongly bound nucleus. The typical timescales, temperatures, and densities involved in nuclear fusion reactions are indicated in Table 10.1. In fact only the most massive stars will evolve to the iron/nickel stage. Less massive stars have smaller gravitational potential energies, and hence (because of the virial theorem) lower thermal energies and lower core temperatures. Stars of $M < 5M_{\text{sun}}$, for example, cease further thermonuclear fusion when they have attained carbon/oxygen cores.

Table 10.1 Nuclear fusion timescales for a star of 25 solar masses (after Rolfs and Rodney 1988)

Fusion of:	Time to complete	Core temperature	Core density (kg m^{-3})
H	7×10^6 yr	6×10^7 K	5×10^4
He	5×10^5 yr	2×10^8 K	7×10^5
C	600 yr	9×10^8 K	2×10^8
Ne	1 yr	1.7×10^9 K	4×10^9
O	0.5 yr	2.3×10^9 K	1×10^{10}
Si	1 day	4.1×10^9 K	3×10^{10}

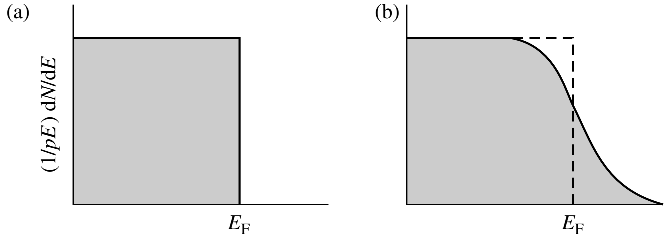


Fig. 10.5 Distribution of electron energies (a) for an electron gas at absolute zero temperature, with all levels filled up to the Fermi energy; and (b) for an electron gas at finite, low temperature, where electrons begin to spill over into states above the Fermi energy.

10.6 Electron degeneracy pressure and stellar stability

At high densities, such as occur in stellar cores at an advanced stage of the evolutionary path, a new form of pressure, in addition to gas pressure and radiation pressure, called *electron degeneracy pressure*, becomes important. To investigate the role of this degeneracy pressure, consider a gas of electrons at absolute zero temperature. The electrons will fall into quantum states of the lowest possible energy, and for this reason the gas is said to be degenerate. The Pauli exclusion principle applies to such identical fermions, so that each quantum state can be occupied by one electron only. At zero temperature, the energy is minimized if all the states are occupied, up to some maximum energy called the Fermi energy ε_F, and all states of energy $\varepsilon > \varepsilon_F$ are unoccupied (see Fig. 10.5). The corresponding momentum is called the Fermi momentum p_F. For values of temperature T above zero, not all these quantum states are filled and the energy spectrum extends above the Fermi energy. Ultimately, when $kT >> \varepsilon_F$, the energy distribution reverts to the Fermi–Dirac distribution described by equation (5.56).

Going back to the case of the completely degenerate electron gas of Fig. 10.5(a), the number of electrons in a physical volume V with momentum $p < p_F$ will be

$$N = g_e V \int \frac{4\pi p^2 \mathrm{d}p}{h^3} = g_e V \frac{4\pi p_F^3}{(3h)^3} \tag{10.19}$$

where $g_e = 2$ is the number of spin substates of each electron, and $4\pi p^2 \mathrm{d}p/h^3$ is the number of states in phase space per unit volume. The number density of electrons will be $n = N/V$ so that

$$p_{\mathrm{F}} = h\left[\frac{3n}{(8\pi)}\right]^{1/3} \tag{10.20}$$

If the electrons are *non-relativistic*, that is, $p_{\mathrm{F}} \ll m_e c$, their kinetic energy is $p^2/2m_e$ and the kinetic energy density will be

$$\frac{E_{\mathrm{NR}}}{V} = \int 8\pi p^2\left(\frac{p^2}{2m_e}\right)\frac{\mathrm{d}p}{h^3} = \frac{8\pi p_{\mathrm{F}}^5}{10m_e h^3} \tag{10.21}$$

Using (5.24b), the degeneracy pressure of the electron gas will therefore be

$$\begin{aligned} P_{\mathrm{NR}} &= \left(\frac{2}{3}\right)\left(\frac{E_{\mathrm{NR}}}{V}\right) = \frac{8\pi p_{\mathrm{F}}^5}{\left(15m_e h^3\right)} \\ &= \left[\left(\frac{3}{8\pi}\right)^{2/3} \times \left(\frac{h^2}{5m_e}\right)\right] n^{5/3} \end{aligned} \tag{10.22}$$

However, if the electrons are mainly *relativistic*, that is, if $p_{\mathrm{F}} \gg m_e c$, the electron energy $\varepsilon \approx pc$ and hence

$$\frac{E_R}{V} = \int \frac{8\pi p^2 \mathrm{d}p\,(pc)}{h^3} = \frac{2\pi p_{\mathrm{F}}^4 c}{h^3} \tag{10.23}$$

The corresponding pressure from (5.24a) becomes

$$P_R = \left(\frac{1}{3}\right)\frac{E_R}{V} = \left[\left(\frac{hc}{4}\right)\cdot\left(\frac{3}{8\pi}\right)^{1/3}\right] n^{4/3} \tag{10.24}$$

Note that, in both relativistic and non-relativistic cases, the pressure increases with the electron density, but more rapidly in the non-relativistic case. This turns out to be crucially important in discussing stability of stars in their final evolutionary stages.

Let us now consider the effects of the gravitational pressure. The gravitational energy of a star or of a stellar core of mass M, radius R, volume V, and an assumed uniform mass density ρ is

$$E_{\mathrm{grav}} = \left(\frac{3}{5}\right)\frac{\mathrm{G}M^2}{R}$$

The volume-averaged gravitational pressure is (see Example 10.2 below)

$$\begin{aligned} P_{\mathrm{grav}} &= \frac{E_{\mathrm{grav}}}{3V} = \left(\frac{\mathrm{G}}{5}\right)\left(\frac{4\pi}{3}\right)^{1/3} M^{2/3}\rho^{4/3} \\ &= \left(\frac{\mathrm{G}}{5}\right)\left(\frac{4\pi}{3}\right)^{1/3} M^{2/3}\left(\frac{m_{\mathrm{P}}A}{Z}\right)^{4/3} n^{4/3} \end{aligned} \tag{10.25}$$

In the second line, the mass density ρ of the material of the stellar core containing the degenerate electron gas has been expressed in terms of the electron number density n, where Z and A are the atomic and mass numbers of the nuclei of the core, respectively , and m_P is the nucleon mass, so that $n = (Z/A)\,\rho/m_{\mathrm{P}}$.

Example 10.2 *Show that, in stellar equilibrium, the volume-averaged gas pressure is one third of the gravitational energy density.*

Let $P(r)$ be the gas pressure at radius r, and consider a spherical shell of density $\rho(r)$ and thickness dr. The outward force on the shell due to the gas pressure is

$$4\pi r^2 \left[P(r) - \left\{ P(r) + \left(\frac{\mathrm{d}P}{\mathrm{d}r} \right) \mathrm{d}r \right\} \right]$$

where $\mathrm{d}P/\mathrm{d}r$ is clearly negative. The inward gravitational force on the shell is $\mathrm{G}M\ (< r)\,\mathrm{d}m/r^2$ where $\mathrm{d}m = 4\pi r^2 \rho(r) dr$ is the mass of the shell. Equating these forces, multiplying through by r and integrating from $r = 0$ to $r = R$, the stellar radius, we get

$$-\int 4\pi r^3 \left(\frac{\mathrm{d}P}{\mathrm{d}r} \right) \mathrm{d}r = -\int \frac{\mathrm{G}M\ (< r)\,\mathrm{d}m}{r}$$

If we integrate the left-hand side by parts and assume $P = 0$ at $r = R$, we get

$$3 \int 4\pi r^2 \mathrm{d}r\, P(r) = 3 \int P(r)\, \mathrm{d}V = -\int \frac{\mathrm{G}M\ (< r)\,\mathrm{d}m}{r} = E_{\mathrm{grav}}$$

where dV denotes the volume of the spherical shell. The second term in this equation represents the gas pressure integrated over volume, while the third is the total gravitational potential energy. Hence we find for the volume-averaged pressure

$$\langle P \rangle = -\left(\frac{1}{3} \right) \frac{E_{\mathrm{grav}}}{V} \qquad (10.26)$$

Provided that the electrons are non-relativistic, the degeneracy pressure (10.22), varying as $n^{5/3}$, will win over the gravitational pressure, varying as $n^{4/3}$. So the star or stellar core will be stable against contraction, since any increase in density will increase the outward electron pressure relative to the inward gravitational pressure. In the case of relativistic degeneracy, however, both pressures have the same density dependence, that is, varying as $n^{4/3}$, so such a core, with mass exceeding the so-called Chandrasekhar limit, is not stable and a suitable trigger can send it into a state of collapse.

The condition $p_{\mathrm{F}} < m_e c$ that the electron momentum is non-relativistic implies from (10.20) that the average distance $n^{-1/3}$ between electrons exceeds the electron Compton wavelength:

$$n^{-1/3} > \left(\frac{3}{8\pi} \right)^{1/3} \left(\frac{h}{m_e c} \right) \sim 0.5 \left(\frac{h}{m_e c} \right) \qquad (10.27)$$

The number density of nucleons is An/Z, so that the critical density for stability will be

$$\rho_0 = \left(\frac{8\pi m_{\mathrm{P}} A}{3Z} \right) \left(\frac{m_e c}{h} \right)^3 \qquad (10.28)$$

Equating the gravitational pressure (10.25) with the non-relativistic degeneracy pressure (10.22) gives a value for the density at which the two are equal:

$$\rho = \left(\frac{4m_e^3 \mathrm{G}^3 M^2}{h^6}\right)\left(\frac{Am_\mathrm{P}}{\mathrm{Z}}\right)^5\left(\frac{4\pi}{3}\right)^3 \tag{10.29}$$

Identifying this with the critical density (10.28), and inserting $G = \hbar c/M_\mathrm{PL}^2$, we get a value for the maximum mass of a stellar core which is stable against collapse:

$$\begin{aligned} M_\mathrm{Ch} &= \left(\frac{3\sqrt{\pi}}{2}\right)\left(\frac{Z}{A}\right)^2\left(\frac{M_\mathrm{PL}}{m_\mathrm{P}}\right)^3 m_\mathrm{P} \\ &= 4.9\left(\frac{Z}{A}\right)^2 M_\mathrm{sun} \qquad (10.30) \\ &\approx 1.2\, M_\mathrm{sun} \end{aligned}$$

where a core of helium or heavier elements, with $Z/A = 1/2$, has been assumed. A more sophisticated stellar model, in which the density varies with radius, yields the more realistic value

$$M_\mathrm{Ch} = 1.4\, M_\mathrm{sun} \tag{10.31}$$

The quantity M_Ch is called the *Chandrasekhar mass*, after one of the physicists who discussed the stability of white dwarfs (see Stoner (1929, 1930) and Chandrasekhar (1931)).

10.7 White dwarf stars

Stars of relatively low mass, such as the Sun in our galaxy, after passing through the hydrogen- and helium-burning phases, will form cores of carbon and oxygen. The higher temperature of the core will lead to helium being burned in a spherical shell surrounding the core, and the stellar envelope will expand by a huge factor and eventually escape to form a planetary nebula surrounding the star. For stars in the solar mass range, the central temperature will not increase enough to lead to carbon burning, so that after the helium is finished, there is no longer any fusion energy source. Nevertheless, the star, providing its mass is less than the Chandrasekhar mass, is saved from catastrophic collapse because of the electron degeneracy pressure in the core. Such a star, bereft of its envelope and slowly cooling off, is known as a *white dwarf*.

All main sequence stars of about one solar mass will end up eventually as white dwarfs. However, these stars are limited to a fairly narrow mass range. The upper limit is determined by the Chandrasekhar mass of 1.4 M_sun, but there is also a lower limit of approximately 0.25 M_sun, since it turns out that the evolution of stars and their emergence as white dwarfs of masses below this limit would be on a timescale much longer than the present age of the universe (see caption to Fig. 10.3). If they are partners of binary systems, the masses of white dwarfs can be measured, and those observed appear to satisfy the above limits.

Example 10.3 *Using the criterion of non-relativistic electron degeneracy, estimate the radius and density of a white dwarf star of one solar mass.*

From (10.29) the density needed to balance the gravitational pressure with non-relativistic degeneracy pressure is

$$\rho = \left(\frac{4m_e^3 G^3 M^2}{h^6}\right) \times \left(\frac{Am_p}{Z}\right)^5 \times \left(\frac{4\pi}{3}\right)^3$$

Inserting $\rho = 3M/4\pi R^3$ gives $R = 7 \times 10^6$m $= 0.01\ R_{\text{sun}}$ for $M = M_S, A/Z = 2$. Note that $R \propto 1/M^{1/3}$, that is, the radius of a white dwarf *decreases* as the mass increases. The average density in the case chosen is clearly 10^6 times that of the Sun, that is, about 2×10^9 kg m^{-3}.

The typical radius of a white dwarf can be estimated as in the above example. It is of order 1% of the solar radius, corresponding to the fact that the average density is of order 10^6 times the mean solar density. In the above discussion, we have treated the density of the white dwarf as uniform, but as is clear from Example 10.2, both the pressure and the density must increase towards the centre of the star. For a white dwarf of about one solar mass, the central density is calculated to be in the region of 10^{11} kg m^{-3}. Since white dwarfs, as the name implies, emit white light, they have surface temperatures of the same order as that of the Sun, so that with about 100 times smaller radius, their luminosities are of order 10^{-3} of the solar luminosity. This guarantees that, even with no nuclear energy source, white dwarfs can continue shining for billions of years.

10.8 Stellar collapse: type II supernovae

A star of mass $M > 10\ M_{\text{sun}}$ is massive enough that it can evolve through all the stages of stellar fusion, ending up eventually with an iron core, produced by silicon burning at a temperature of order 4×10^9 K, as sketched in Fig. 10.4. As more silicon is burned in a shell surrounding the iron, both the mass of the iron core and its temperature will increase, until eventually the core mass itself exceeds the Chandrasekhar limit (10.31). The core is then unstable and is driven into collapse by two triggering mechanisms: photodisintegration of iron nuclei by thermal photons, and the conversion of electrons to neutrinos by inverse beta decay. The result is a supernova explosion, with a rate of order one per century in spiral galaxies such as the Milky Way.

As the collapse proceeds, some of the gravitational energy released goes into heating up of the core to well above 10^{10} K, which is a mean thermal photon energy above 2.5 MeV, so that a fraction of the photons can cause photodisintegration of the iron nuclei into alpha particles (helium nuclei). With enough photons it is clear that iron can be systematically broken down completely into helium, thus reversing the effects of all the fusion processes since the main sequence *pp* cycle:

$$\gamma + {}^{56}\text{Fe} \longleftrightarrow 13\,{}^{4}\text{He} + 4n \tag{10.32}$$

This equation indicates that an equilibrium between iron and helium nuclei is set up, in which the balance swings to the right-hand side as the core temperature increases. Of course, the absorption of energy in the above endothermic process

(145 MeV for the complete photodisintegration of each iron nucleus into alphas) further speeds up gravitational collapse, the core heats still further and the helium nuclei themselves undergo photodisintegration:

$$\gamma + {}^4\mathrm{He} \longleftrightarrow 2p + 2n \tag{10.33}$$

The collapse also heralds the onset of *neutronization*, in which the electrons from the degenerate 'sea' convert free or bound protons to neutrons via inverse beta decay:

$$\mathrm{e}^- + p \longleftrightarrow n + \nu_e \tag{10.34}$$

with 0.8 MeV threshold. When the radius of the core has collapsed by an order of magnitude and the density has reached the vicinity of 10^{12} kg m^{-3}, the Fermi momentum of the electrons is, from (10.20)

$$p_{\mathrm{F}}c = hc\left[\frac{3Z\rho}{(8\pi A m_{\mathrm{P}})}\right]^{1/3} \sim 4 \text{ MeV} \tag{10.35}$$

so that an electron of energy near the Fermi energy can trigger the above reaction, or an equivalent inverse beta decay in iron:

$$\mathrm{e}^- + {}^{56}\mathrm{Fe} \longleftrightarrow {}^{56}\mathrm{Mn} + \nu_e \tag{10.36}$$

with a threshold of 3.7 MeV. As more and more electrons are converted to neutrinos in these processes, the degeneracy pressure of the electrons will decrease steadily and the collapse will then be virtually unopposed. The free-fall time of the collapse will be given by (8.34):

$$t_{\mathrm{FF}} = \left(\frac{3\pi}{32\,\mathrm{G}\rho}\right)^{1/2} \sim 0.1 \text{ s} \tag{10.37}$$

Eventually, the collapse is halted, as the density reaches nuclear density. The gravitational pressure is then opposed by the degeneracy pressure of the (non-relativistic) nucleons. The core contains iron nuclei, electrons, and protons as well as a preponderance of neutrons (hence the name neutron star). Very roughly, we can treat the collapsed core as a gigantic nucleus of neutrons, so that we expect the radius to be of the order of

$$R = r_0 A^{1/3} \tag{10.38}$$

where $r_0 = 1.2$ fm $= 1.2 \times 10^{-15}$ m is the unit nuclear radius from measured nuclear radii. For a core mass $M = 1.5\,M_{\mathrm{sun}}$, the mass number $A = M/m_{\mathrm{P}} \sim 1.9 \times 10^{57}$ and the radius $R \sim 15$ km, for a nuclear density $\rho_N = 3m_{\mathrm{P}}/4\pi r_0^3 \sim 2 \times 10^{17}$ kg m^{-3}. The repulsive nuclear force at short distances resists further compression, and it is estimated that as soon as the density exceeds nuclear density by about a factor of 2–3 times, the collapse will be brought to an abrupt halt, and the core material will 'bounce', producing an outgoing pressure wave which develops into a supersonic shock wave which traverses the infalling material of the envelope and finally—it is still not absolutely clear how—gives rise to the spectacular optical phenomenon of a supernova explosion. Such an event, resulting from the collapse of a massive star, is known as a Type II supernova.

As the initial collapse proceeds, the reactions (10.34) and (10.36) will result in emission of neutrinos. In particular a short, few millisecond burst of $10^{56} - 10^{57}$ neutrinos ν_e will accompany the outgoing shock wave. They will have energies of a few MeVs and account for up to 5% of the total gravitational energy released. However, as soon as the core density exceeds about 10^{15} kg m^{-3}, it becomes effectively opaque, even to neutrinos, and they become trapped in the contracting material.

The total gravitational energy released in the collapse to a neutron star of 1.5 solar masses and uniform density will be (taking the above values of A and r_0):

$$\begin{aligned} E_{\text{grav}} &= \left(\frac{3}{5}\right) \mathrm{G}m_{\mathrm{P}}^2 \frac{A^{5/3}}{r_0} \\ &= 3.0 \times 10^{46}\ \mathrm{J} \qquad (10.39) \\ &= 1.8 \times 10^{59}\ \mathrm{MeV} \end{aligned}$$

which amounts therefore to about 100 MeV per nucleon. Note that this implies that an original uncontracted mass of 1.55 M_{sun} of nucleons will result in a neutron star of mass only 1.4 M_{sun}. This energy release is much larger than the energy required to disintegrate the iron into its constituent nucleons (the binding energy per nucleon in iron is 8 MeV) or to convert protons and electrons to neutrons and neutrinos (0.8 MeV per nucleon). The huge amount of energy released, however, remains temporarily locked in the core, which enters a 'thermal phase' in which photons, electron–positron pairs and neutrino–antineutrino pairs, together with the neutrons and a few protons and heavier nuclei reach thermal equilibrium. All flavours of neutrino and antineutrino will be generated in this thermal phase:

$$\gamma \longleftrightarrow \mathrm{e}^+ + \mathrm{e}^- \longleftrightarrow \nu_i + \bar{\nu}_i \qquad (10.40)$$

where $i = e, \mu, \tau$. The mean free path for neutrinos in the core material depends on both charged and neutral current scattering by nucleons, electrons, and nuclei. As an indication we consider the charged current scattering (10.41) of ν_e by neutrons. The cross-section is of order $G_{\mathrm{F}}^2 p_f^2$ from (1.18) and (1.27), where $p_{\mathrm{f}} = E \sim (E_\nu - Q)$ is the neutrino energy above the (negative) threshold ($Q = -0.8$ MeV)—see Problem 10.7. In detail the cross-section is

$$\sigma\left(\nu_e + n \to p + \mathrm{e}^-\right) = \left(\frac{G_{\mathrm{F}}^2}{\pi}\right)\left[1 + 3g_{\mathrm{A}}^2\right]E^2 = 0.94 \times 10^{-43} E^2\ \mathrm{cm}^2 \qquad (10.41)$$

where $g_{\mathrm{A}} = 1.26$ is an axial–vector coupling constant and E is in MeV. For a typical nuclear density of $\rho = 2 \times 10^{17}$ kg m^{-3}, the neutrino mean free path or diffusion length would be of order $\lambda = 1/\sigma N_0 \rho \sim 900/E^2$ m, which for a typical value of $E = 20$ MeV, means $\lambda \sim 2$ m only. This process is just one of neutrino absorption. For neutral current scattering of neutrinos (the only option for muon- and tau-neutrinos), the mean free path would be several times bigger, of order 5–10 m.

In each neutral current scattering process (analogously to the case of photon diffusion through the Sun in Example 10.1 above), the neutrino will emerge in an arbitrary direction, so that after N successive scatters, this 'random walk' will carry the neutrino a root mean square straight line distance of $\lambda N^{1/2}$. Identifying this with the core radius R, we obtain a diffusion time from the central region to the surface of $t \sim R^2/\lambda c \sim 0.1 - 1$ s. Since neutrinos are the only particles which are able to escape, the 100 MeV gravitational energy release per nucleon is divided among the six flavours of neutrino/antineutrino, and detailed computer simulations indeed indicate that neutrinos and antineutrinos of all flavours are emitted from the core in comparable numbers over a period of 0.1–10 s, with average energy $\sim$ 15 MeV and with an approximately Fermi–Dirac distribution as in (5.56). They are emitted from a so-called neutrinosphere within a few metres of the surface. Neutrinos account for 99% of the total gravitational energy released in (10.39). The spectacular optical display of a Type II supernova explosion accounts for only 1% of the total energy release.

10.9 Neutrinos from SN 1987A

Figure 10.6 shows a photograph of the supernova SN1987A in the Large Magellanic Cloud, a mini-galaxy about 60 kpc from the Milky Way. It is famous because it was the first supernova from which interactions of the emitted neutrinos have been observed, in fact simultaneously in the large Kamiokande (Hirata *et al.* 1987) and IMB (Bionta *et al.* 1987) water Cerenkov detectors (with masses of 2 kton and 7 kton respectively) and in the smaller (0.2 kton) Baksan liquid scintillator detector (Alekseev *et al.* 1987). (All these detectors were originally designed to search for proton decay.) The neutrino pulse was actually detected about 7 h before the optical signal became detectable.

The principal reactions that could lead to detection of supernova neutrinos in a water detector are as follows:

$$\bar{\nu}_e + p \to n + \mathrm{e}^+ \tag{10.42a}$$

$$\nu + \mathrm{e}^- \to \nu + \mathrm{e}^- \tag{10.42b}$$

$$\bar{\nu} + \mathrm{e}^- \to \bar{\nu} + \mathrm{e}^- \tag{10.42c}$$

The secondary electrons or positrons from these reactions have relativistic velocities and part of their energy loss in traversing the water appears in the form of Cerenkov light (see Section 9.6) which is detected by an array of photomultipliers, as in Fig. 4.7.

The first reaction (10.42a) has a threshold of $Q = 1.8$ MeV and a cross-section rising as the square of the neutrino energy, as in (1.23), with a value of 10^{-41} cm^2 per proton at $E_\nu = 10$ MeV. The angular distribution of the secondary lepton is almost isotropic. The second and third reactions go via both neutral and charged current channels for electron-neutrinos/antineutrinos, and via neutral currents only for muon and tauon neutrinos/antineutrinos. Although not negligible, the summed cross-section for these reactions (which vary as E_ν) is only 10^{-43} cm^2 per electron at 10 MeV. So, although in water there are five electrons for every free proton, the event rate for scattering off electrons is an order of magnitude less than that for the first reaction. Moreover in (10.42a) all the particles will

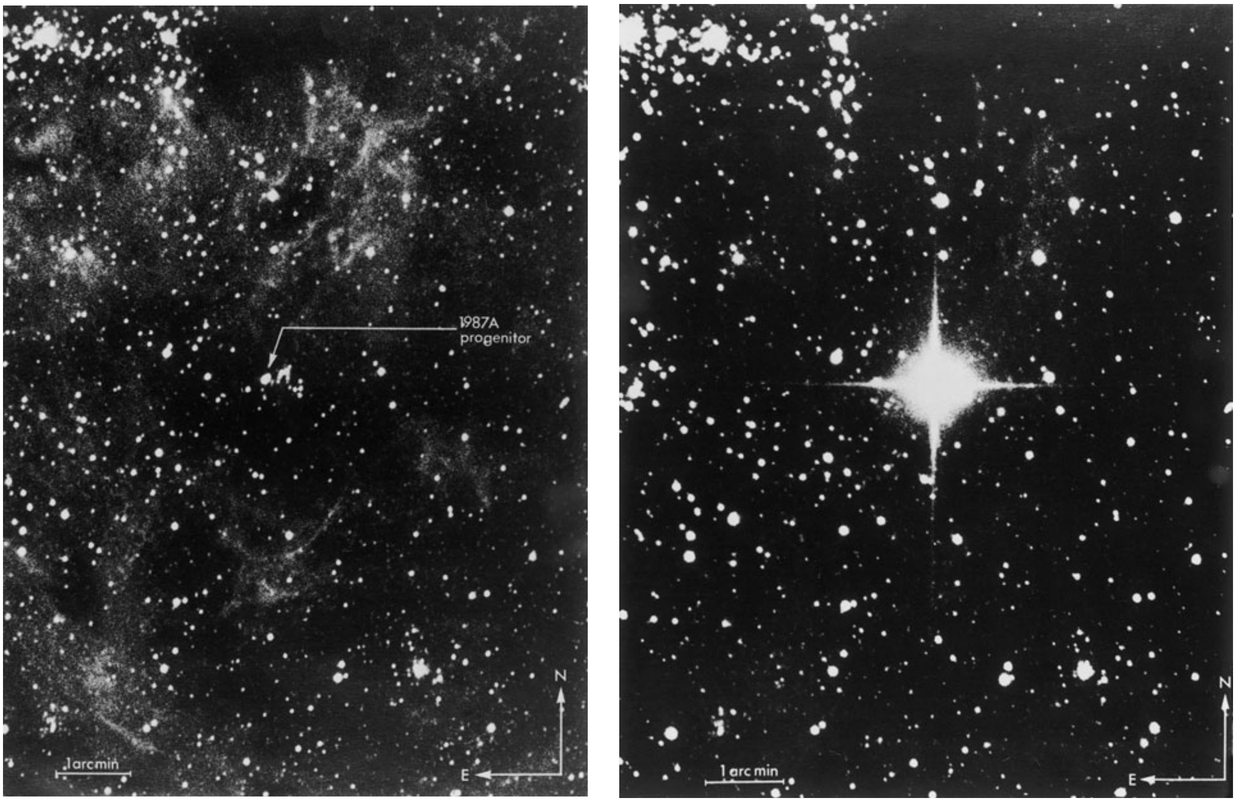


Fig. 10.6 The SN 1987A supernova. The stellar field in the Large Magellanic Cloud before (left) and two days after (right) the supernova explosion. Although such a supernova is for some time the brightest object in the local galaxy, the light emitted is only about 1% of the total energy released. The rest is accounted for by neutrinos. In this particular case, the progenitor star was a blue giant of about 20 solar masses. No neutron star (pulsar) has been detected as a remnant.

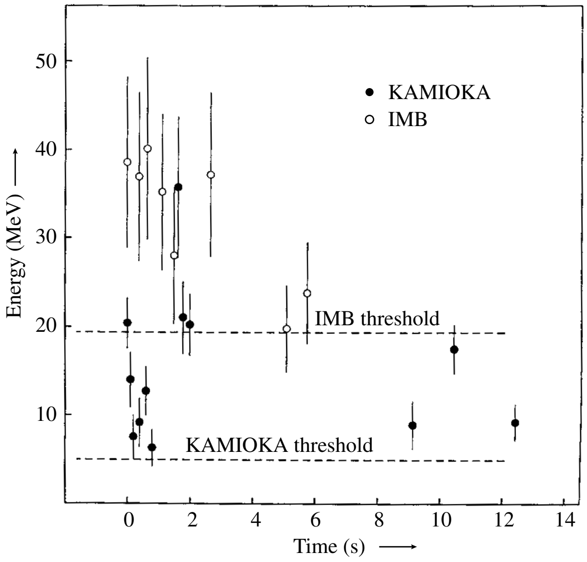


Fig. 10.7 The energies of the IMB and Kamiokande water Cerenkov events plotted against arrival time. The effective threshold energies for the two detectors were 20 and 6 MeV respectively.

have comparable momenta, so that on account of its mass, the proton kinetic energy will be very small and the secondary positron will receive most of the energy ($E_e = E_\nu - 1.8\,\text{MeV}$); while in (10.42b) and (10.42c) the charged lepton receives typically half the incident energy.

The event rates recorded, together with the known distance to the supernova (60 kpc) could be used to compute the total energy flux in neutrinos and antineutrinos, assuming that the total is six times that for ν_e alone. Both data sets, when account is taken of detection thresholds, are consistent with a mean temperature of $kT \sim 5\,\text{MeV}$ and thus an average neutrino energy at production of 3.15 kT appropriate to a relativistic Fermi–Dirac distribution. The integrated neutrino luminosity thus calculated from the event rates was

$$\begin{aligned} L &\approx 3 \times 10^{46}\,\text{J} \\ &\approx 2 \times 10^{59}\,\text{MeV} \end{aligned} \tag{10.43}$$

with an estimated uncertainty of a factor two, and thus in excellent agreement with the prediction (10.39).

It perhaps needs to be emphasized that the neutrino flux from a Type II supernova is indeed prodigious. Altogether some 10^{58} neutrinos were emitted from SN1987A and even at the Earth, about 170,000 light years distant, the flux was over 10^{10} neutrinos cm^{-2}.

The recording of neutrinos from SN1987A gave some information on neutrino properties. The fact that they survived a 170,000 year journey without attenuation testifies to their stability. Since the neutrino pulse lasted less than 10 s, the transit time of neutrinos of different energies was the same within 1 part in 5×10^{11}. The time of arrival on the Earth, t_E will be given in terms of the emission time from the supernova, t_SN, its distance L, and the neutrino mass m and energy E by

$$t_\text{E} = t_\text{SN} + \left(\frac{L}{c}\right)\left[1 + \left(\frac{m^2c^4}{2E^2}\right)\right]$$

for $m^2 \ll E^2$. For two events with different energies E_1 and E_2 the time difference will be given by

$$\Delta t = |\Delta t_\text{E} - \Delta t_\text{SN}| = \left(\frac{Lm^2c^4}{2c}\right)\left[\frac{1}{E_1^2} - \frac{1}{E_2^2}\right] \tag{10.44}$$

If we take as typical values $E_1 = 10\,\text{MeV}$, $E_2 = 20\,\text{MeV}$ and $\Delta t < 10\,\text{s}$, this equation gives $m < 20\,\text{eV}$. A more refined calculation does not result in a better limit.

It is of interest to remark here that the neutrino burst may be instrumental in helping the shock wave to develop the spectacular optical display of a supernova. Early computer models suggested that the outward moving shock might stall as it met with the infalling matter from outside the core, and produced disintegration of this material into its constituent nucleons. However, in at least some of the simulations, when account was taken of the interactions of the neutrinos with the envelope material, the transfer of only 1% of the total neutrino energy was found to be enough to keep the shock wave moving.

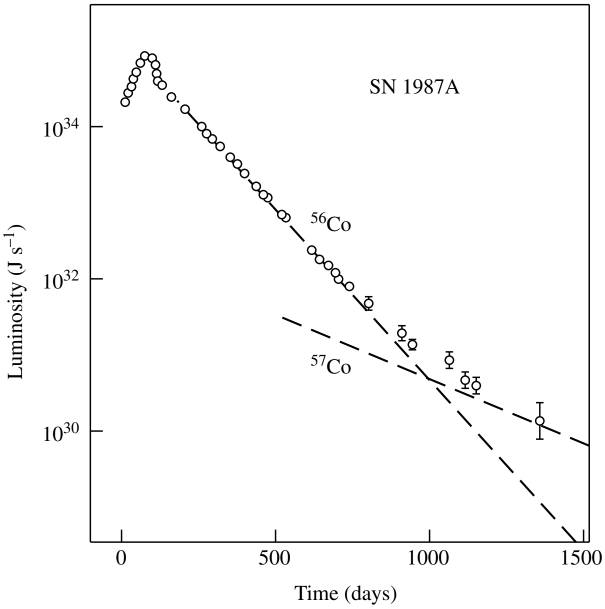


Fig. 10.8 The light curve of supernova SN 1987A. After the initial outburst, the luminosity fell rapidly over the first 100 days, being dominated by the beta decay of ^{56}Ni to ^{56}Co, with a mean lifetime $\tau = 9$ days. From time $t = 100$ to $t = 500$ days, the energy release was dominated by the beta decay of ^{56}Co to ^{56}Fe, with $\tau = 111$ days. Beyond $t = 1000$ days, the important decay is of ^{57}Co to ^{57}Fe ($\tau = 391$ days) as well as that of other long-lived isotopes. Most of the heavy nuclei would have been produced in rapid absorption reactions of neutrons with the material of the infalling envelope. Interestingly enough, no neutron star has been detected following this particular supernova. (After Suntzeff *et al.* 1992.)

So it seems possible that neutrinos of all flavours—ν_e, ν_μ, and ν_τ—interacting *via* both neutral- and charged current processes, still play a vital part in cosmic events, while of course the corresponding charged μ and τ leptons disappeared by decay within microseconds of the Big Bang, and we are left with only the electrons. We may also remark here that such supernovae perform a unique role in the production of very heavy elements, since they are sources of the very intense neutron fluxes which build up the nuclei in the later part of the Periodic Table, *via* rapid neutron capture chains. So it is worth bearing in mind that the iodine in your thyroid and the barium in your bones probably owe their existence to the fact that there are three flavours of neutrino and antineutrino, with both neutral and charged current couplings.

As stated above, only about 1% of the total supernova energy appears in the form of light output (although this is enough to dominate for a time the luminosity of the host galaxy). The light curve, at least for the first 3 years, is approximately exponential, being dominated by the decay of radioactive ^{56}Co, with a mean lifetime of 111 days (see Fig. 10.8).

10.10 Neutron stars and pulsars

The rump left behind after a supernova explosion is usually a neutron star, which contains neutrons, protons, electrons, and heavier nuclei, but with neutrons predominating. In the free state, a neutron undergoes decay with a mean lifetime of 887 ± 2 s, so we must consider the equilibrium in the reversible reaction

$$n \longleftrightarrow p + e^- + \bar{\nu}_e + 0.8 \text{ MeV} \tag{10.45}$$

In the neutron star, the decay of the neutron will be prevented as a result of the Pauli principle, provided that all the quantum states that can be reached by the electron and the proton from the decay are already filled. To a very good approximation, if we neglect the Q-value in the decay, this condition is satisfied when the Fermi energies of the degenerate neutrons and electrons are equal, that is, when $\varepsilon_F(n) = \varepsilon_F(e)$, so that the forward and backward reactions are in equilibrium. We know from (10.39) that the neutrons and protons will be non-relativistic, while the electrons will be ultra-relativistic, so that $p_F(e) \ll p_F(n)$. Then it is clear from (10.20) that the electron number density will be much smaller than that of the neutrons, as shown in the numerical Example 10.4.

Example 10.4 *Estimate the ratio of numbers of electrons (and protons) to neutrons needed to prevent neutron decay in a neutron star of density* $\rho = 2 \times 10^{17}$ *kg m*$^{-3}$.

Assuming that practically all of the nucleons are neutrons, their number density will be $n_n = 1.2 \times 10^{44}$ m^{-3}, and their Fermi momentum from (10.20) will be

$$p_F\,(n)\,c = hc\left[\frac{3n_n}{8\pi}\right]^{1/3} = 300\,\text{MeV}$$

while their (non-relativistic) Fermi energy will be

$$\varepsilon_F\,(n) = \frac{[p_F\,(n)\,c]^2}{2M_n c^2} = 48\text{ MeV}$$

To prevent neutron decay, the Fermi momentum and energy of the (relativistic) electrons therefore need only be of order

$$p_F\,(e)\,c = \varepsilon_F\,(e) = \varepsilon_F\,(n) = 48\text{ MeV}$$

and the electron number density, proportional to the cube of the Fermi momentum, will be $n_e = [48/300]^3\, n_n \sim 0.004\, n_n$. Obviously, $n_p = n_e$ by charge conservation. So a small proportion, less than 1%, of electrons and protons is sufficient to prevent neutron decay, and the equilibrium in (10.45) is very much to the left.

[*Note*: Here we have for simplicity neglected the small effect of the protons. If we include them, the equilibrium condition becomes $\varepsilon_F(n) = \varepsilon_F(e) + \varepsilon_F(p)$. It is left as an exercise to solve the ensuing quadratic equation, and show that the effect of the protons is to reduce the above electron concentration by 7%].

Although the early theory of neutron stars was developed shortly after Chadwick discovered the neutron in 1932, major experimental interest had to await the discovery of *pulsars* by Hewish *et al.* (1968). Pulsars are rapidly rotating neutron stars which emit radiation at short and extremely regular intervals, much like a rotating lighthouse beam which crosses the line of sight of an observer with a regular frequency. As indicated below, pulsars possess enormously strong magnetic fields, which are believed to be responsible for acceleration of cosmic ray protons and nuclei. More than 1000 pulsars are known, with rotational periods ranging from 1.5 ms to 8.5 s. Only about 1% of pulsars can be associated with past supernova remnants, since over millions of years the neutron stars have drifted away from the remnant nebulae. For a few young pulsars like that in the Crab, the nebula is still associated. This most famous example of a pulsar has a period of 33 ms, and is the remnant of the AD 1054 supernova recorded by the Chinese.

In addition to pulsars like that in the Crab which emit radio waves, about 200 *X-ray pulsars* are known. These are neutron stars which are members of binary star systems. Matter accretes from the massive companion star on to the magnetic poles of the neutron star, creating the X-ray emission ('aurora'). The X-rays are pulsed with the rotational frequency of the neutron star.

X-ray bursters are associated with neutron stars which have light main sequence companion stars. Hydrogen is accreted on to the very hot neutron star surface, and after some time, it reaches a density and temperature leading to ignition in a thermonuclear explosion lasting several seconds. The process is repeated as more material is accreted. *γ-Ray bursters* have already been mentioned in Section 9.11. They are associated with the most violent events in the universe, releasing an estimated 10^{46} J in γ-rays, about the same amount of energy as in a Type II supernova explosion. They are possibly produced as a result of neutron star binaries merging to form black holes.

The maximum angular frequency ω of a pulsar will be given by the requirement that the outward centrifugal force on the surface material should not exceed the inward gravitational attraction, that is,

$$\omega^2 R < \frac{\mathrm{G}M}{R^2} \tag{10.46}$$

Inserting the values $R \sim 15\,\mathrm{km}$, $M \sim 1.5\,M_{\mathrm{Sun}}$ gives for the minimum period $\tau = 2\pi/\omega \sim 1\,\mathrm{ms}$, and indeed no neutron stars are observed with shorter periods than this. The very high frequencies observed for many pulsars result because much of the angular momentum of the original giant stars is retained through later evolutionary stages and the rotational frequency is enormously increased because of the dramatic contraction to the neutron star size.

The pulsar radiation itself is ascribed to the existence of a rotating magnetic dipole inclined at an angle θ to the axis of rotation. For a dipole of strength μ the electromagnetic power radiated is proportional to the square of the radial acceleration, that is, to ω^4:

$$P \propto \mu\,\omega^4 \sin^2\theta \tag{10.47}$$

The magnetic field at the surface of a pulsar is of order $10^8 T$, this high value resulting from the trapping and concentration of magnetic flux by the highly conducting plasma during stellar collapse, the field increasing inversely as the square of the radius. The energy lost through the emission of the radiation results in a small deceleration of the pulsar. If I is the moment of inertia of the pulsar, the rotational energy is $1/2I\omega^2$, so that the rate of change of rotational energy or power emitted is

$$P = I\omega\left(\frac{\mathrm{d}\omega}{\mathrm{d}t}\right) \propto \omega^4$$

so that

$$\frac{\mathrm{d}\omega}{\mathrm{d}t} = -A\omega^3$$

Observations on the Crab pulsar indicate that at the present time, $d\omega/dt = -2.4 \times 10^{-9}$ while $\omega = 190\,\mathrm{s}^{-1}$. If the initial angular velocity is ω_i, the time t for which it has been spinning to reach the present value of ω is

$$t = \left(\frac{1}{2A}\right)\left[\frac{1}{\omega^2} - \frac{1}{\omega_i^2}\right] < \frac{1}{\left(2A\omega^2\right)} = \frac{1}{2}\frac{\omega}{(d\omega/dt)} = 1255 \text{ yr.}$$

in agreement with its origin in AD 1054.

10.11 Black holes

Neutrons play a similar role in supporting a neutron star as degenerate electrons do in supporting a white dwarf. The limit to which the degenerate neutron gas can do this is analogous to the Chandrasekhar limit for electron degeneracy in white dwarfs. If we forget about the strong nuclear interactions and general relativity effects in high gravitational fields, we can apply (10.29) with the substitution $A/Z = 1$, so that the limit (10.30) therefore becomes

$$M_\mathrm{max} \sim 4.9\, M_\mathrm{sun} \tag{10.48}$$

For neutron stars with masses $M > M_\mathrm{max}$, the degenerate neutron gas becomes relativistic, and gravitational collapse is inevitable. However, strong interactions between the neutrons will tend to make the neutron star matter more incompressible and increase the maximum mass. On the other hand, the gravitational binding energy of a neutron star is comparable with its mass energy (see (10.39)), so that non-linear gravitational effects associated with the mass energy of the field itself should be included and this will tend to decrease the maximum mass of the neutron star. So (10.48) should only be taken as an indication that the critical mass of neutron stars is a few solar masses.

The fate of a neutron star which undergoes gravitational collapse is a *black hole*. Not all black holes are formed in this way. For example, the giant low-metallicity Wolf–Rayet stars are believed to collapse to black holes at the end of the carbon burning phase. The crucial property associated with a black hole is the Schwarzschild radius (2.23) for an object of mass M, given by the formula

$$r_\mathrm{s} = \frac{2\mathrm{G}M}{c^2} \tag{10.49}$$

This result was derived by Schwarzschild, who obtained an exact solution to Einstein's field equations of general relativity, for the specific case of a gravitational field due to a large static mass M. As explained in Chapter 2, it happens that it can also be found using special relativity and the equivalence principle, or by equating the radial escape velocity from a point mass M to the velocity of light. As an example of (10.49), the Schwarzschild radius of a star of mass $M = 5M_\mathrm{sun}$ is $r_\mathrm{s} = 15\,\mathrm{km}$. Equation (10.49) implies that when the physical radius of a collapsed star falls inside the Schwarzschild radius, there are no light paths (geodesics) to the outside world. Photons from the star cannot escape its gravitational field and the star becomes black to an outside observer.

To understand how can this be, let us apply the special theory of relativity by comparing a time interval dt on a stationary clock in a remote inertial frame with that, $\mathrm{d}t'$, on an identical clock stationary in an inertial frame which has velocity v relative to the first frame, and is instantaneously co-moving with the surface of the collapsing star. Then $\mathrm{d}t'^2 = \mathrm{d}t^2\left(1 - v^2/c^2\right)$. Thus as $v \to c$, $\mathrm{d}t' \to 0$ and to an observer in the remote frame, the star appears 'frozen' in time. Equally, the wavelength of light from a star collapsing inside its Schwarzschild radius undergoes a redshift according to the relation $\lambda' = \lambda/\left(1 - \left(v^2/c^2\right)\right)^{1/2}$, and the redshift tends to infinity as $v \to c$. Since the quantum energy of the radiation $h\nu = hc/\lambda \to 0$ as $\lambda \to \infty$, the energy emitted from the star also tends to zero. These phenomena are what would be recorded by an external observer. An observer *within* the Schwarzschild radius would, however, record lots of activity, but would not be able to communicate with the outside world.

Black holes are inevitable consequences of Einstein's general theory of relativity—even if Einstein could not bring himself to believe in their existence. From the general theory it follows that everything possessing energy and momentum, including photons, will be deflected by a gravitational field and will be 'turned around' if the field is strong enough. The experimental evidence for the existence of black holes is quite convincing. It rests, for example, on observation of binary systems in which the motion of the visible star implies the existence of a compact invisible companion with mass $M > M_{\max}$. Such systems are observed as compact X-ray sources, the X-rays being produced as mass from the visible star flows into the black hole. The first candidate for a black hole was the X-ray source Cygnus X-1, with $M = 3.4\,M_{\text{sun}}$. Another candidate was V404 Cygni, which included a compact object of $M > 6\,M_{\text{sun}}$. Recent investigations show that intense amounts of X-rays are emitted from the centres of nearly all galaxies. The conclusion is that the X-rays are emitted by very hot gas flowing into a massive central black hole, typically with a mass of $10^6 - 10^8$ solar masses. For example, the Milky Way possesses a black hole at its centre of 3×10^6 solar masses, identified with the X-ray source Sagittarius A^* (see Problem 10.7).

It is generally considered that some of the most violent events in the universe, such as γ-ray bursts (see Section 9.11) originate from the very small proportion ($\sim 1\%$) of galactic nuclei which are 'active'—the so-called AGNs. These are associated with massive black holes at the centres of galaxies which are still very active in absorbing large amounts of nearby material—stars, gas, and dust—as distinct from most galactic nuclei, with black holes which have gone through that stage and are now relatively passive. Recently, as mentioned in Chapter 9, it has been found that charged cosmic ray primaries with energies exceeding 6×10^{19} eV are correlated in direction with AGNs, presumably because they are accelerated in the neighbourhood of black holes.

At the other extreme, primordial 'mini black holes' might have been created in the early universe, but if so equation (10.51) below suggests that they would have evaporated long ago if their masses were much below 10^{12} kg.

10.12 Hawking radiation from black holes

When quantum fluctuations are brought into the picture, it turns out that in the very strong gravitational fields around them, black holes are actually able to emit (thermal) radiation, as proved by Hawking in 1974. The Hawking temperature for a black hole of mass M is given by

$$kT_{\mathrm{H}} = \frac{\hbar c^3}{(8\pi \mathrm{G}M)} \tag{10.50}$$

For example, for $M = 5\,M_{\mathrm{sun}}, T_{\mathrm{H}} \sim 1.23 \times 10^{-8}$ K. Note that as the black hole loses energy and mass, it gets hotter, and thus a black hole will eventually evaporate and disappear. The lifetime can be calculated from the rate of energy loss from the surface:

$$\frac{\mathrm{d}\left(Mc^2\right)}{\mathrm{d}t} = 4\pi r_{\mathrm{s}}^2 \sigma T_{\mathrm{H}}^4$$

where $\sigma = \pi^2 k^4 / \left(60\hbar^3 c^2\right)$ is the Stefan constant. Substituting from (10.49) and (10.50) and integrating, one obtains for the lifetime

$$\begin{aligned}\tau_{\mathrm{BH}} &= \text{constant} \times \frac{\mathrm{G}^2 M^3}{\left(\hbar c^4\right)} \\ &\sim 10^{67} \left(\frac{M}{M_{\mathrm{sun}}}\right)^3 \text{ yrs}\end{aligned} \tag{10.51}$$

Thus the time for a black hole of a typical astronomical mass to evaporate is far longer than the age of the universe.

The origin of the Hawking radiation and the form of (10.50) can be made plausible by the following simple argument. Suppose that, as a result of a quantum fluctuation at a radial distance r just at the Schwarzschild radius of a black hole of mass M, a virtual $\mathrm{e}^+\mathrm{e}^-$ pair is temporarily created (see Fig. 10.9). Under normal conditions, such a pair would quickly annihilate, but in the presence of an enormously strong gravitational field, any small separation can lead to a tidal force which is enough to convert at least one member into a real state. If the pair has total energy E it can, according to the uncertainty principle, exist for a time $\Delta t \sim \hbar/E$. In this time the two particles can separate by a maximum radial distance $\Delta r \sim c \cdot \Delta t \sim c\hbar/E$. The difference of the gravitational field strengths at the positions of the two particles is then $\left(2GM/r^3\right)\Delta r$ and the difference ΔF of the gravitational forces upon them is this quantity multiplied by the effective mass, E/c^2, so that the tidal force is

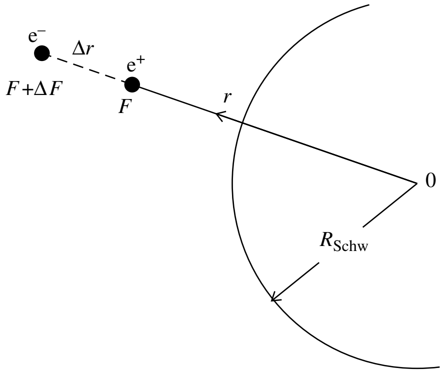


Fig. 10.9 Creation of an electron–positron pair just outside the Schwarzschild radius of a black hole.

$\Delta F \sim \left(GM/r^3\right)\hbar/c$. For the gravitational field to be able to create the pair, one requires $\Delta F \Delta r > E$, that is, $E < \hbar\left(GM/r^3\right)^{1/2}$. The largest value of E will be for the minimum value of r, namely $r \sim r_s$, giving the order-of-magnitude condition

$$E \sim \frac{\hbar c^3}{GM}$$

agreeing with (10.50) up to numerical constants. In the presence of the strong gravitational field, the pair can be separated fast enough that one of them gets well outside of the Schwarzschild radius and escapes as a real particle, while the other is sucked back into the black hole.

10.13 Summary

- Stars form from protostars consisting of vast clouds of hydrogen in gravitational collapse, which contract until the core temperature reaches $kT \sim 1$ keV, when thermonuclear fusion of hydrogen to helium commences and the star attains hydrostatic equilibrium. This hydrogen fusion on the main sequence of the Herzsprung–Russell diagram continues typically for billions of years.
- When the hydrogen fuel is exhausted, fusion of helium to carbon and oxygen takes place, at a higher core temperature and on a much shorter timescale. The star becomes a red giant with a bloated envelope.
- If the stellar mass is of the order of a solar mass or less, the consumption of helium marks the end of the fusion cycle and of nuclear energy release, and the star cools off slowly as a white dwarf with a degenerate electron core.
- In more massive stars, the core temperature becomes high enough for fusion to continue, with production of heavier elements up to nickel and iron.
- If the mass of the iron core exceeds the Chandrasekhar mass of 1.4 solar masses, it is inherently unstable and suffers catastrophic collapse with formation of a very compact neutron star of nuclear density. About 10% of the mass energy of the star is emitted in the form of a burst of neutrinos, and as a shock wave which gives rise to the optical display of a Type II supernova.
- If the core mass is much larger, around 4 to 6 solar masses, the neutron star is unstable and undergoes further collapse to a black hole. Binary systems with a black hole as one partner are compact and very intense sources of X-rays, emitted as matter flows into the black hole from the companion.
- Black holes can decay by emission of Hawking radiation, which is a manifestation of quantum fluctuations near the Schwarzschild radius of a black hole. Intense sources of X-rays from the centres of many galaxies are attributed to emission from very hot gas flowing into massive (10^8 solar mass) black holes at the galactic centre. Such black holes may be identified with active galactic nuclei, associated with intense γ-ray bursts as described in Chapter 9.

Problems

More challenging problems are marked with an asterisk.

(10.1) Estimate the maximum rotational frequency of a white dwarf star, assuming that it has a mass equal to a solar mass and radius of 1% of the solar radius.

(10.2) Calculate the luminosity (in watts) of the Sun, given that its surface temperature is 5780 K and its radius is 7×10^8 m. Solar energy is provided by fusion of helium from hydrogen. If 5% of the hydrogen in the Sun has so far been converted to helium, estimate the age of the Sun, assuming the luminosity to have been constant.

*(10.3) Find the maximum mass for a body containing normal atomic matter (density 10^4 kg m^{-3}) which does not require electron degeneracy pressure in order to maintain its stability against gravitational collapse. What object might such a body represent?

(10.4) The slow-down of the Crab Pulsar is assumed to be due to emission of dipole radiation as a result of its rotating magnetic field as indicated in (10.47). However, one could also ask if the slow-down could be due to gravitational quadrupole radiation, varying with rotational frequency as ω^6 as in (9.28). Show that the observed values of $\omega = 190\,\mathrm{s}^{-1}$ and $\mathrm{d}\omega/\mathrm{d}t = -2.4 \times 10^{-9}$ would then be inconsistent with its known age.

*(10.5) Calculate the mass of a black hole with a Schwarzschild radius equal to the particle horizon distance for a universe of age t_0. If the universe had a density equal to the critical density, at what value of t_0 would the mass of the universe be equal to that of the black hole?

(10.6) Estimate the radius and mass of a black hole with a lifetime equal to that of the universe.

(10.7) Show that, in the reaction $\bar{\nu}_e + p \to \mathrm{e}^+ + n$, the cross-section is of order $G_\mathrm{F}^2 p_f^2$ where the CMS momentum in the final state $p_\mathrm{f} \approx E_\nu - Q$. Here E_ν is the antineutrino energy, assumed to be small compared with the nucleon rest-mass but large compared with the electron rest-mass, which can be neglected, and Q is the threshold energy for the reaction.

(10.8) A star has been observed (in the infrared) in orbit about a massive unseen object (black hole) at the centre of our galaxy, identified with the compact radio and X-ray source Sagittarius A (see *Physics Today*, February 2003 for reference). The elliptic orbit has a period of 15 years and eccentricity of 0.87, and the closest distance of approach (perigee) is estimated to be 17 light hours. If necessary referring to a text on celestial mechanics, calculate the mass of the black hole and the orbital velocity of the star at perigee.

Table of physical constants

A

The following table is taken from the 'Physical Constants' Table of the Particle Data Group, published in the *European Physical Journal* **C15**, 1 (2000). Constants in the table below are quoted only to three figures of decimals.

Symbol	Name	Value
c	velocity of light (in vacuum)	$2.998 \times 10^{8}\ \mathrm{m\,s^{-1}}$
$\hbar$	Planck's constant/2π	$1.055 \times 10^{-34}\ \mathrm{J\,s} = 6.582 \times 10^{-22}\ \mathrm{MeV\,s}$
$\hbar c$		$0.197\,\mathrm{GeV\,fm} = 3.16 \times 10^{-26}\ \mathrm{J\,m}$
e	electron charge	$1.602 \times 10^{-19}\ \mathrm{C}$
m_{e}	electron mass	$0.511\ \mathrm{MeV}/c^2 = 9.109 \times 10^{-31}\ \mathrm{kg}$
m_{p}	proton mass	$0.938\ \mathrm{GeV}/c^2 = 1.672 \times 10^{-27}\ \mathrm{kg} = 1836 m_{\mathrm{e}}$
m_{n}	neutron mass	$0.939\ \mathrm{GeV}/c^2$
$m_{\mathrm{n}} - m_{\mathrm{p}}$	neutron–proton mass difference	$1.293\ \mathrm{MeV}/c^2$
μ_0	permeability of free space	$4\pi \times 10^{-7}\ \mathrm{N\,A^{-2}}$
$\varepsilon_0 = 1/\mu_0 c^2$	permittivity of free space	$8.854 \times 10^{-12}\ \mathrm{F\,m^{-1}}$
$\alpha = e^2/4\pi\varepsilon_0\hbar c$	fine structure constant	1/137.036
$r_{\mathrm{e}} = e^2/4\pi\varepsilon_0 m_{\mathrm{e}} c^2$	classical electron radius	$2.818 \times 10^{-15}\ \mathrm{m} = 2.818\ \mathrm{fm}$
$a_\infty = 4\pi\varepsilon_0\hbar^2/m_{\mathrm{e}} e^2$	$= r_{\mathrm{e}}/\alpha^2 =$ Bohr radius	$0.529 \times 10^{-10}\ \mathrm{m}$
$\bar{\lambda}_{\mathrm{c}} = \hbar/m_{\mathrm{e}} c = r_{\mathrm{e}}/\alpha$	= reduced Compton wavelength of electron	$3.861 \times 10^{-13}\ \mathrm{m}$
$\sigma_{\mathrm{T}} = 8\pi r_{\mathrm{e}}^2/3$	Thomson cross section	$0.665 \times 10^{-28}\ \mathrm{m^2} = 0.665\ \mathrm{b}$
$\mu_{\mathrm{B}} = e\hbar/2m_{\mathrm{e}}$	Bohr magneton	$5.788 \times 10^{-11}\ \mathrm{MeV\,T^{-1}}$
$\mu_{\mathrm{N}} = e\hbar/2m_{\mathrm{p}}$	nuclear magneton	$3.152 \times 10^{-14}\ \mathrm{MeV\,T^{-1}}$
$\omega/B = e/m_{\mathrm{e}}$	cyclotron frequency of electron	$1.759 \times 10^{11}\ \mathrm{rad\,s^{-1}\,T^{-1}}$
N_{A}	Avogadro's number	$6.022 \times 10^{23}\ \mathrm{mol^{-1}}$
k	Boltzmann's constant	$1.381 \times 10^{-23}\ \mathrm{J\,K^{-1}} = 8.617 \times 10^{-11}\ \mathrm{MeV\,K^{-1}}$
$\sigma = \pi^2 k^4/60\hbar^3 c^2$	Stefan's constant	$5.670 \times 10^{-8}\ \mathrm{W\,m^{-2}\,K^{-4}}$
$G_{\mathrm{F}}/(\hbar c)^3$	Fermi coupling constant	$1.166 \times 10^{-5}\ \mathrm{GeV^{-2}}$
$\sin^2\theta_{\mathrm{W}}$	weak mixing parameter	0.2312
M_{W}	W-boson mass	$80.42\ \mathrm{GeV}/c^2$
M_{Z}	Z-boson mass	$91.19\ \mathrm{GeV}/c^2$
G	gravitational constant	$6.673 \times 10^{-11}\ \mathrm{m^3\,kg^{-1}\,s^{-2}}$
au	astronomical unit = mean Earth–Sun distance	$1.496 \times 10^{11}\ \mathrm{m}$
$M_{\mathrm{PL}} = (\hbar c/G)^{1/2}$	Planck mass	$1.221 \times 10^{19}\ \mathrm{GeV}/c^2 = 2.177 \times 10^{-8}\ \mathrm{kg}$

Symbol	Name	Value
pc	parsec	3.086×10^{16} m = 3.262 ly
$M_\odot$	solar mass	1.989×10^{30} kg
$R_\odot$	solar radius (equator)	6.961×10^{8} m
$L_\odot$	solar luminosity	3.85×10^{26} W
$M_\oplus$	Earth mass	5.975×10^{24} kg
$R_\oplus$	Earth radius (equator)	6.378×10^{6} m
H_0	Hubble expansion rate	72 ± 3 km s^{-1} Mpc^{-1}
T_0	CMBR temperature	2.725 ± 0.001 K
t_0	age of Universe	14 ± 2 Gyr

Conversion factors
1 eV = 1.602×10^{-19} J; 1 eV/c^2 = 1.782×10^{-36} kg
kT at 300 K = 1/38.681 = 0.02585 eV
1 erg = 10^{-7} J; 1 dyne = 10^{-5} N; 1 cal = 4.18 J; 0 °C = 273.15 K
1 atmosphere = 760 Torr = 101 325 Pa = 1013 gm cm^{-2}
1 barn = 10^{-28} m^2; π = 3.141592; e = 2.718281828
pc = $0.3B\rho$ = momentum in GeV of singly-charged particle with radius of curvature ρ in metres in a magnetic field of B tesla.

Yukawa theory and the boson propagator

B

The propagator term involved in the exchange of virtual bosons in the interactions between elementary particles arises in the theory of quantum exchange first proposed by Yukawa in 1935. Yukawa was seeking to describe the short-range nature of the potential between neutrons and protons in the atomic nucleus. He started with the relativistic relation between total energy E, three-momentum p and rest-mass m as in (1.1):

$$E^2 = p^2c^2 + m^2c^4 \tag{B.1}$$

We now substitute the coordinate operators $E_{\text{op}} = -i\hbar\partial/\partial t$ and $p_{\text{op}} = -i\hbar\nabla$, which will yield the expectation values of energy and momentum when applied to the wavefunction of a particle, so that the above equation then becomes (dividing through by $-\hbar^2c^2$):

$$\left(\frac{1}{c^2}\right)\frac{\partial^2\psi}{\partial t^2} = \nabla^2\psi - \left(\frac{m^2c^2}{\hbar^2}\right)\psi \tag{B.2}$$

called the *Klein–Gordon wave equation* describing the propagation of a free, spinless particle of mass m. If we insert $m = 0$, (B.2) becomes the familiar wave equation describing the propagation of an electromagnetic wave with velocity c, with ψ interpreted either as the wave amplitude of the associated photons, or as the electromagnetic potential $U(\mathbf{r})$. For a static, radially symmetric potential, we drop the time-dependent term so that (B.2) assumes the form

$$\nabla^2 U(r) \equiv \left(\frac{1}{r^2}\right)\frac{\partial}{\partial r}\left(\frac{r^2\partial U}{\partial r}\right) = \left(\frac{m^2c^2}{\hbar^2}\right)\psi \tag{B.3}$$

As can be verified by substitution, integration of this expression yields the solution

$$U(r) = \left(\frac{g_0}{4\pi r}\right)\exp\left(-\frac{r}{R}\right) : R = \frac{\hbar}{mc} \tag{B.4}$$

In this expression, g_{o} is a constant of integration. In the electromagnetic case $m = 0$ and the static potential is $U(r) = Q/4\pi r$ where Q is the electric charge at the origin. Hence Yukawa interpreted g_{o} as the 'strong nuclear charge'. Inserting for R the known range of nuclear forces of about 1.4 fm, one obtains $mc^2 = \hbar c/R \sim 150$ MeV. The pion, first observed in cosmic rays in 1947 was a particle of zero spin and just this mass. However, the interpretation of nuclear forces in terms of heavy quantum exchange turns out to be much more complicated than Yukawa had envisaged 70 years ago—for example, it involves spin-dependent potentials. Nor is the pion a fundamental boson but just the lightest quark–antiquark combination. Nevertheless, Yukawa's theory pointed to a fundamental relation between the range of the interaction (B.4) between two elementary particles and the mass of the associated exchange quantum, which is just as valid today as it was years ago.

Let us consider a particle of incident momentum $\mathbf{p}_i$ being scattered with momentum $\mathbf{p}_f$ by the potential $U(\mathbf{r})$ provided by a massive source, in which case no energy is transferred and the numerical value of the momentum p of incident and scattered particle are the same. The particle will be deflected through some angle θ and receive a momentum transfer $\mathbf{q} = \mathbf{p}_i - \mathbf{p}_f \left(= 2p \sin [\theta/2]\right)$. The amplitude$f(\mathbf{q})$ for scattering will be the Fourier transform of the potential $U(\mathbf{r})$, in exactly the same way that the angular distribution of light diffracted by an obstacle in classical optics is the Fourier transform of the spatial extent of the obstacle. If g represents the coupling of the particle to the potential, we can write

$$f(\mathbf{q}) = g \int U(\mathbf{r}) \exp\left(i\,\mathbf{q}\cdot\mathbf{r}\right)\, \mathrm{d}V \tag{B.5}$$

Assuming a central potential $U(\mathbf{r}) = U(r)$ and with $\mathbf{q}\cdot\mathbf{r} = qr\cos\theta$ and $\mathrm{d}V = r^2 \mathrm{d}r\mathrm{d}\phi\sin\theta\mathrm{d}\theta$ where θ and ϕ are polar and azimuthal angles, and introducing the Yukawa potential (B.4) we obtain

$$\begin{aligned} f(\mathbf{q}) &= 2\pi g \iint U(r) \exp(iqr \;\; \cos\theta)\; \mathrm{d}(\cos\theta)\; r^2\, \mathrm{d}r \\ &= gg_0 \int \exp\left(-\frac{r}{R}\right) \left\{ \frac{\left(\exp[iqr] - \exp[-iqr]\right)}{iqr} \right\} r^2\, \mathrm{d}r \\ &= \frac{gg_0}{\left(\mathbf{q}^2 + \left(1/R^2\right)\right)} = \frac{gg_0}{\left(\mathbf{q}^2 + \left(m^2c^4 \big/ \hbar^2\right)\right)} \end{aligned} \tag{B.6}$$

This result is for a massive potential source, where three-momentum but no energy has been exchanged. For an actual scattering process between two particles, the relativistically invariant four-momentum transfer squared will be $q^2 = \Delta E^2 - \Delta\mathbf{p}^2 = \Delta E^2 - \mathbf{q}^2$. So for $\mathbf{q}^2$ in (B.6), holding for $\Delta E = 0$, we should substitute $-q^2$, so that the scattering amplitude becomes, in units $\hbar = c = 1$

$$f(q^2) = \frac{gg_0}{\left[m^2 - q^2\right]} \tag{B.7}$$

Thus the scattering amplitude denoted by $|T_{\mathrm{fi}}|$ in Section 1.8 consists of the product of the couplings of the two particles to the exchanged virtual boson, multiplied by the propagator term, which depends on the four-momentum transferred (where q^2 is always negative) and on the mass of the free boson. All the above expressions are for spinless particles, and additional factors are required when spin is introduced.

Perturbative growth of structure in the early universe

C

We start with the FLRW model described in Chapters 2 and 5, which assumes a completely isotropic and homogeneous distribution of matter and radiation undergoing the Hubble expansion. We assume we are dealing, at least initially, with tiny perturbations and therefore weak gravitational fields. Further the distances involved, although enormous, are assumed to be small compared with the horizon distance *ct* so that the Hubble flow is non-relativistic. Thus we can use Newtonian mechanics based on classical fluid dynamics. There are three basic equations which read as follows:

$$\frac{\partial \rho}{\partial t} + \nabla\cdot(\rho \mathbf{u}) = 0 \tag{C.1}$$

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u} = -\left[\left(\frac{1}{\rho}\right)\nabla P + \nabla\Phi\right] \tag{C.2}$$

$$\nabla^2\Phi = 4\pi G\rho \tag{C.3}$$

In these equations, ρ is the fluid density, $\mathbf{u}$ the velocity of fluid flow, P the pressure and Φ is the gravitational potential. The first equation (C.1) is the *equation of continuity* expressing the fact that the rate of decrease of fluid density with time is just equal to the divergence of the fluid flow, that is the mass of fluid flowing out of the volume in unit time. Equation (C.2) is *Euler's equation*. It states that the force ∇P on a volume element is equal to the rate of change of momentum of that element, that is $\rho d\mathbf{u}/dt = -(\nabla P + \rho\nabla\Phi)$ if the gravitational pressure is included. The total rate of change of velocity, applying to a particular element of the fluid, is made up of two parts: the partial derivative $\partial\mathbf{u}/\partial t$, measuring the change in fluid velocity at a particular spatial coordinate, plus a term due to the fact that the liquid element is in motion and in time dt it has travelled a distance d$\mathbf{r} = \mathbf{u}$dt. Thus

$$\begin{aligned}\mathrm{d}\mathbf{u} &= \mathrm{d}t\left(\frac{\partial \mathbf{u}}{\partial t}\right) + \left[\mathrm{d}\mathbf{x}\left(\frac{\partial \mathbf{u}}{\partial x}\right) + \mathrm{d}\mathbf{y}\left(\frac{\partial \mathbf{u}}{\partial y}\right) + \mathrm{d}\mathbf{z}\left(\frac{\partial \mathbf{u}}{\partial z}\right)\right] \\ &= \mathrm{d}t\left(\frac{\partial \mathbf{u}}{\partial t}\right) + (\mathrm{d}\mathbf{r}\cdot\nabla)\mathbf{u} \end{aligned} \tag{C.4}$$

and the result (C.2) follows upon dividing by dt. The third equation (C.3) is *Poisson's equation* for the gravitational potential in terms of G and the density.

In the absence of any perturbations in density, the above equations have the following solutions:

$$
\begin{aligned}
\rho(t) &= \frac{\rho_0}{[R(t)]^3} \\
\mathbf{u}(t,\mathbf{r}) &= \left[\frac{\dot{R}(t)}{R(t)}\right]\mathbf{r} \qquad\qquad (\mathrm{C.5}) \\
\Phi(t,r) &= \frac{2\pi \mathrm{G}\rho r^2}{3}
\end{aligned}
$$

The first expresses the dependence of density on the expansion parameter $R(t)$. In equation (C.1), assuming that we are dealing with a homogeneous universe, we have $\nabla\rho = 0$, while $\rho\ \nabla\cdot\mathbf{u} = \rho\left(\dot{R}/R\right)\ \nabla\cdot\mathbf{r} = 3\,\rho\left(\dot{R}/R\right)$: the result follows (see footnote to (C.8)). The second is the equation for Hubble flow $\mathbf{u}(t,\mathbf{r}) = H(t)\mathbf{r}$, and the third follows from integration of (C.3), using $\nabla^2\Phi = \left(1/r^2\right)\left[\partial\left(r^2\partial\Phi\big/\partial r\right)/\partial r\right]$ in spherical coordinates.

Now we suppose that perturbations in the values of **u** and ρ occur. It turns out to be easier to discuss the developments in a coordinate frame co-moving with the Hubble expansion. In the following, **r** denotes a position coordinate measured by a 'stationary' observer (i.e. one not moving with the Hubble flow), and **x** that in the co-moving frame. Then $\mathbf{x} = \mathbf{r}/R(t)$. The velocity of a fluid particle defined above as **u** in the stationary frame is then

$$\mathbf{u} = \frac{d\mathbf{r}}{dt} = \mathbf{x}\dot{R} + \mathbf{v} \qquad (\mathrm{C.6})$$

The first term on the right measures the velocity arising from the Hubble flow, and the extra term **v** (where $\mathbf{v} \ll \mathbf{u}$) is the so-called 'peculiar velocity' of the particle relative to the general expansion. In the absence of a perturbation, this would of course be zero. The perturbation in density ρ is denoted $\Delta\rho \ll \rho$ and the fractional change, called the 'density contrast', is denoted $\delta = \Delta\rho/\rho$. A gradient in the stationary system is denoted $\boldsymbol{\nabla}_{\mathrm{s}}$ to distinguish it from that in the co-moving frame, called $\boldsymbol{\nabla}_{\mathrm{c}}$, where

$$\boldsymbol{\nabla}_{\mathrm{c}} = R\boldsymbol{\nabla}_{\mathrm{s}} \qquad (\mathrm{C.7})$$

Finally, time derivatives of any function, say F, in the two systems will be related by

$$\left(\frac{\partial F}{\partial t}\right)_{\mathrm{s}} = \left(\frac{\partial F}{\partial t}\right)_{\mathrm{c}} - \dot{R}\mathbf{x}\cdot\frac{(\boldsymbol{\nabla}_{\mathrm{c}}F)}{R} \qquad (\mathrm{C.8})$$

where the velocity of the stationary frame is—$\dot{R}\mathbf{x}$ with respect to the co-moving frame. With these definitions the continuity equation (C.1) will read[1]

$$\left[\frac{\partial}{\partial t} - \left(\frac{\dot{R}}{R}\right)\mathbf{x}\cdot\boldsymbol{\nabla}_{\mathrm{c}}\right]\rho(1+\delta) + \frac{\rho}{R}\boldsymbol{\nabla}_{\mathrm{c}}\cdot[(1+\delta)(\dot{R}\mathbf{x}+\mathbf{v})] = 0 \qquad (\mathrm{C.9})$$

In evaluating this expression, recall that $\boldsymbol{\nabla}\rho = 0$ in a homogeneous universe. Further, if the pressure is small, that is we are dealing with non-relativistic matter in our cosmic fluid, $\rho \propto 1/R^3$, so $\partial\rho/\partial t = -3\rho\dot{R}/R$. The quantity

[1] The following relations are useful in evaluating (C.3) and (C.5):

$$
\begin{aligned}
\nabla\cdot\mathbf{r} &= \left(\mathbf{i}\frac{\partial}{\partial x} + \mathbf{j}\frac{\partial}{\partial y} + \mathbf{k}\frac{\partial}{\partial z}\right) \\
&\quad \times(\mathbf{i}x + \mathbf{j}y + \mathbf{k}z) = 3 \\
(\mathbf{v}\cdot\nabla)\mathbf{x} &= (\mathbf{i}v_{\mathrm{x}} + \mathbf{j}v_{\mathrm{y}} + \mathbf{k}v_{\mathrm{z}}) \\
&\quad \times\left(\mathbf{i}\frac{\partial}{\partial x} + \mathbf{j}\frac{\partial}{\partial y} + \mathbf{k}\frac{\partial}{\partial z}\right) \\
&\quad \times[\mathbf{i}x + \mathbf{j}y + \mathbf{k}z] \\
&= \left(v_x\frac{\partial}{\partial x} + v_{\mathrm{y}}\frac{\partial}{\partial y} + v_z\frac{\partial}{\partial z}\right) \\
&\quad \times[\mathbf{i}x + \mathbf{j}y + \mathbf{k}z] \\
&= (\mathbf{i}v_{\mathrm{x}} + \mathbf{j}v_{\mathrm{y}} + \mathbf{k}v_{\mathrm{z}}) = \mathbf{v}
\end{aligned}
$$

$\nabla_c\cdot\mathbf{x} = 3$, so that $(\rho/R)\ \nabla_c\cdot\dot{R}\ \mathbf{x} = +3\rho\dot{R}/R$. Finally, second order terms such as the product $v\delta$ can be neglected. The equation then reads

$$\left(\frac{\partial\delta}{\partial t}\right) + \frac{\nabla_c\cdot\mathbf{v}}{\mathbf{R}} = 0 \tag{C.10}$$

The Euler equation (C.2) becomes

$$\begin{aligned}&\left[\frac{\partial}{\partial t} - \left(\frac{\dot{R}}{R}\right)\mathbf{x}\cdot\nabla_c\right](R\mathbf{x}+\mathbf{v}) + (R\mathbf{x}+\mathbf{v})\cdot\nabla_c\frac{(R\mathbf{x}+\mathbf{v})}{R}\\ &\quad = -\frac{[\nabla_c\Phi + (\partial P/\partial\rho)\nabla_c(1+\delta)]}{R}\end{aligned} \tag{C.11}$$

Subtracting the equation for the unperturbed system and again neglecting second order perturbative terms such as $\mathbf{v}\cdot\nabla\mathbf{v}$ gives

$$\frac{\partial\mathbf{v}}{\partial t} + \mathbf{v}\frac{\dot{R}}{R} + \frac{\nabla_c\phi}{R} + v_s^2\frac{\nabla_c\delta}{R} = 0 \tag{C.12}$$

where ϕ (assumed to be $\ll\ \Phi$) is the gravitational potential due to the perturbation and $\partial P/\partial\rho = v_s^2$ determines the speed of sound in the fluid. The Poisson equation gives $\nabla^2\phi = 4\pi\mathrm{G}\rho\delta$, so that subtracting the time derivative of (C.10) from the divergence of (C.12) gives

$$\frac{\partial^2\delta}{\partial t^2} + 2\frac{\dot{R}}{R}\left(\frac{\partial\delta}{\partial t}\right) - 4\pi G\delta\rho - \frac{v_s^2(\nabla^2\delta)}{R^2} = 0 \tag{C.13}$$

The final step is to express the spatial dependence of the pressure and density perturbation as a superposition of plane waves of wavenumbers k, that is of the form

$$\delta(x,t) = \sum\delta_k(t)\exp(\mathrm{i}\mathbf{k}\cdot\mathbf{x}) \tag{C.14}$$

so that for a particular wavenumber k, (C.7) becomes

$$\frac{\mathrm{d}^2\delta}{\mathrm{d}t^2} + 2\left(\frac{\dot{R}}{R}\right)\frac{\mathrm{d}\delta}{\mathrm{d}t} = \left(4\pi G\rho - \frac{k^2v_s^2}{R^2}\right)\delta \tag{C.15}$$

The terms on the right vanish for a value of k corresponding to the Jeans length:

$$\lambda_J = \frac{2\pi R}{k} = v_s\left(\frac{\pi}{G\rho}\right)^{1/2} \tag{C.16}$$

First we note that, if the expansion of the universe is neglected, that is $\dot{R}(t) = 0$, the solution of (C.15) is either periodic or exponential, according to the two possibilities:

i. $\lambda \gg \lambda_J$: if the response time for the pressure wave is large compared with the gravitational infall time, the density contrast *grows exponentially*:

$$\delta \propto \exp\left(\frac{t}{\tau}\right) \quad \text{where } \tau = \left[\frac{1}{(4\pi G\rho)}\right]^{1/2} \tag{C.17}$$

ii. $\lambda \ll \lambda_J$: in this case the solution to (C.15) is of the form

$$\delta \propto \exp(i\omega t) \quad \text{where } \omega = \frac{2\pi v_s}{\lambda} \tag{C.18}$$

so that the density contrast *oscillates as a sound wave*.

C.1 Growth in the matter-dominated era

In the early stages of the Big Bang, the universe is radiation dominated and in that case the velocity of sound is relativistic, with a value $v_s = (\partial P/\partial\rho)^{1/2} = c/\sqrt{3}$—see Table 5.2. This means that, using equation (5.47) with $\rho_r c^2 = \left(3c^2/32\pi G\right)/t^2$, the Jeans length is

$$\lambda_J = c\left[\frac{\pi}{3G\rho_r}\right]^{1/2} = ct\left(\frac{32\pi}{9}\right)^{1/2} \tag{C.19}$$

In this case the Jeans length and the horizon distance are both of order ct, where $t = 1/H$ is the Hubble time (i.e. the time since the start of the Big Bang). Thus growth in this stage of the radiation era may appear less likely (and it is also true that our assumption of classical Newtonian mechanics in Euclidean space might not be valid on such large length scales).

After radiation and matter decouple, that is at a temperature of $kT \sim 0.3$ eV when $t \approx 3 \times 10^5$ year, the electrons and protons combine to form hydrogen atoms and the velocity of sound and hence *the Jeans length will decrease abruptly*, so that growth of inhomogeneities becomes possible. At the above decoupling temperature, $v_s \sim 5 \times 10^3$ m s^{-1} only and thus the Jeans length has decreased by more than 10^4 times.

Let us take the simple case of a matter-dominated universe with $\rho = \rho_c$ and $\Omega = 1$, usually referred to as an Einstein-de Sitter universe. Then from (5.26) and (5.15)

$$\rho = \frac{3H^2}{8\pi G} \quad \text{and} \quad H = \frac{\dot{R}}{R} = \frac{2}{3t}$$

so that (C.15) becomes, assuming $4\pi G\rho \gg \frac{k^2 v_s^2}{R^2}$

$$\frac{d^2\delta}{dt^2} + \frac{4}{3t}\frac{d\delta}{dt} - \frac{2}{3t^2}\delta = 0 \tag{C.20}$$

which has a power law solution of the form

$$\delta = At^{2/3} + Bt^{-1} \tag{C.21}$$

where A and B are constants. The second term describes a contracting mode and is of no interest. The first term describes a mode in which the density contrast grows as a power law. Thus the effect of taking into account the expansion of the universe is to change an exponential growth as in (C.17) to a power law dependence. We note from (5.2) that

$$\frac{\delta_0}{\delta_{\text{dec}}} = \left(\frac{t_0}{t_{\text{dec}}}\right)^{2/3} = \frac{R(0)}{R_{dec}} = (1 + z_{\text{dec}}) \approx 1100 \tag{C.22}$$

where the symbols ‘0’ and ‘dec’ refer to quantities today and at the time of decoupling of matter and radiation, that is when electrons and protons started to combine to form hydrogen atoms. Since today $kT_0 = 0.23$ meV, and at the time of decoupling $kT_{\mathrm{dec}} \approx 0.3$ eV, $(1 + z_{\mathrm{dec}}) = 1100$, as in (5.75). The above equation is a result, based on the assumption of small perturbations, and since we started out with a density contrast of order 10^{-5}, such a big extrapolation may be questionable. Nevertheless, this analysis shows that any small anisotropies at the time of decoupling of matter and radiation will increase in proportion to the scale parameter $R(t)$.

D The MSW mechanism in solar neutrino interactions

We start discussion of the MSW mechanism by writing out the time evolution of the mass eigenstates (4.8), given by the Schroedinger equation $i\mathrm{d}\psi/\mathrm{d}t = E\psi$ for the time dependence of the wavefunction. In matrix form, using (4.9) this appears as:

$$\begin{aligned} i\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix} &= \begin{pmatrix} E_1 & 0 \\ 0 & E_2 \end{pmatrix}\begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix} \\ &= \begin{pmatrix} m_1^2/2p & 0 \\ 0 & m_2^2/2p \end{pmatrix}\begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix} + \begin{pmatrix} p & 0 \\ 0 & p \end{pmatrix}\begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix} \qquad \text{(D.1)} \end{aligned}$$

The term on the extreme right is a constant phase factor which affects ν_1 and ν_2 equally and can therefore be omitted. If we substitute the expression for ν_e and ν_μ in terms of ν_1 and ν_2 in (4.7) we find after a little straightforward algebra that, for vacuum oscillations

$$i\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} \nu_e \\ \nu_\mu \end{pmatrix} = M_V \begin{pmatrix} \nu_e \\ \nu_\mu \end{pmatrix} \qquad \text{(D.2)}$$

where

$$M_V = \left[\frac{\left(m_1^2 + m_2^2\right)}{4p}\right]\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \left(\frac{\Delta m^2}{4p}\right)\begin{pmatrix} -\cos 2\theta & \sin 2\theta \\ \sin 2\theta & \cos 2\theta \end{pmatrix}$$

In interactions with matter, electron neutrinos in the MeV energy range can undergo both charged ($\mathrm{W}^\pm$ exchange) and neutral-current (Z^0 exchange) interactions, while muon- or tau-neutrinos have only the neutral-current option, as their energies are too low to generate the charged lepton. Hence, electron-neutrinos suffer an extra potential affecting the forward scattering amplitude, which leads to a change in the effective mass:

$$\begin{aligned} V_e &= G_\mathrm{F}\sqrt{2}N_e \\ m^2 &= E^2 - p^2 \rightarrow (E + V_e)^2 - p^2 \approx m^2 + 2EV_e \\ \Delta m_m^2 &= 2\sqrt{2}G_\mathrm{F}N_e E \qquad \text{(D.3)} \end{aligned}$$

where N_e is the electron density, $E = pc$ is the neutrino energy, G_F is the Fermi constant and Δm_m^2 is the shift in mass squared. (For antineutrinos, which are the CP transforms of neutrinos, the sign of the potential V_e is reversed.) So in the case of electron neutrinos traversing matter, one should substitute in the vacuum expression for the average mass squared in (D.2):

$$\frac{1}{2}\left(m_1^2+m_2^2\right)\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix} \to \frac{1}{2}\left(m_1^2+m_2^2\right)\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix} + 2\sqrt{2}G_{\mathrm{F}}N_e p\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$$
$$= \left[\frac{1}{2}\left(m_1^2+m_2^2\right)+\sqrt{2}G_{\mathrm{F}}N_e p\right]\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix} + \sqrt{2}G_{\mathrm{F}}N_e p\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}$$

Then the matrix M_M appropriate to matter traversal is modified from the vacuum matrix M_{V} in (D.2) as follows:

$$M_M = \left[\frac{\left(m_1^2+m_2^2\right)}{4p}+\frac{\sqrt{2}G_{\mathrm{F}}N_e}{2}\right]\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix} + \left[\frac{\Delta m^2}{4p}\right]\begin{pmatrix}-\cos 2\theta + A & \sin 2\theta\\ \sin 2\theta & \cos 2\theta - A\end{pmatrix} \quad \text{(D.4)}$$

where $A = 2\sqrt{2}G_{\mathrm{F}}N_e p/\Delta \mathrm{m}^2$. The first term again gives the same phase factor for ν_e and ν_μ and can therefore be omitted. If we denote the mixing angle in the presence of matter as θ_{m} and the mass difference squared in matter as Δm_m^2, the second term in (D.4) can also be written as in (D.2):

$$\left[\frac{\Delta m_m^2}{4p}\right]\begin{pmatrix}-\cos 2\theta_m & \sin 2\theta_m\\ \sin 2\theta_m & \cos 2\theta_m\end{pmatrix} \quad \text{(D.5)}$$

and equating these two gives us immediately

$$\tan 2\theta_m = \frac{\sin 2\theta}{[\cos 2\theta - A]} = \frac{\tan 2\theta}{\left[1-\left(L_{\mathrm{v}}/L_{\mathrm{e}}\right)\sec 2\theta\right]} \quad \text{(D.6)}$$

where the vacuum oscillation length is $L_{\mathrm{v}} = 4\pi p/\Delta m^2$ and the electron interaction length is defined as $L_{\mathrm{e}} = 4\pi/\left(2\sqrt{2}G_{\mathrm{F}}N_{\mathrm{e}}\right)$, so that $A = L_{\mathrm{v}}/L_{\mathrm{e}}$. We note that, irrespective of the value of θ, it is possible for the matter mixing angle to go through a 'resonance' with $\theta_m = \pi/4$, provided that L_{v} is positive and therefore $\Delta m^2 > 0$, that is $m_2 > m_1$. The resonance condition is clearly

$$L_{\mathrm{v}} = L_{\mathrm{e}}\cos 2\theta \qquad \text{or} \qquad N_{\mathrm{e}}(\mathrm{res}) = \frac{\Delta m^2 \cos 2\theta}{\left(2\sqrt{2}G_{\mathrm{F}}p\right)} \quad \text{(D.7)}$$

where $N_{\mathrm{e}}(\mathrm{res})$ denotes the corresponding 'resonant' electron density. For example, the density in the core of the Sun is ρ (core) ~ 100 gm cm^{-3}, or $N_{\mathrm{e}}(\mathrm{core}) \sim 3\times 10^{31}$ m^{-3}, giving $L_{\mathrm{e}} \sim 3\times 10^5$ m (compared with a solar radius of 7×10^8 m). The solar density falls off (roughly exponentially) with radius outside of the core. If at some radius the 'resonance' condition is fulfilled, electron-neutrinos could be transformed partly or entirely to muon– or tauon–neutrinos, even if the vacuum mixing angle is very small. Neutrinos will *always* pass through the resonance region if the critical electron density is less than that in the solar core, that is if the energy exceeds a minimum value

$$E_{\min} = \frac{\Delta m^2 \cos 2\theta}{\left[2\sqrt{2} G_{\mathrm{F}} N_{\mathrm{e}}\,(\text{core})\right]} \tag{D.8}$$

This explains the larger suppression observed in Table 9.1 for the higher energy boron-8 neutrinos.

For an intuitive picture of the MSW effect, let us, for simplicity, assume that the vacuum mixing angle is small. Then an electron–neutrino starts out in the solar core, predominantly in what in a vacuum would be called the ν_1 eigenstate of mass m_1, but the extra weak potential increases the effective mass in the appropriate region of electron density to the value m_2, (refer to (4.6) for $\theta_{\mathrm{m}} = 45°$) which in a vacuum would be identified as consisting principally of the ν_μ flavour eigenstate. If the solar density changes fairly slowly with radius, this mass eigenstate will pass out of the Sun without further changes, and in vacuum it will be identified with the muon–neutrino eigenstate. In general, however, there will only be partial flavour conversion (see Fig. D.1).

One of the most crucial aspects of the data is the observed energy spectrum of electrons from the boron-8 neutrinos in the SUPER-K and SNO experiments, which indicates an almost constant suppression factor over the range 6–14 MeV. The fits to the data indicate a large vacuum mixing angle θ_{12} (originally, there were two solutions, one with a large and one with a small mixing angle, but the latter is now excluded).

The atmospheric and solar neutrino results are shown in Fig. 4.3, where the fitted values of $\tan^2\theta$ are plotted against Δm^2, and the closed shaded areas represent the allowed regions for the parameters. The atmospheric results have been combined with data from long baseline accelerator experiments, as

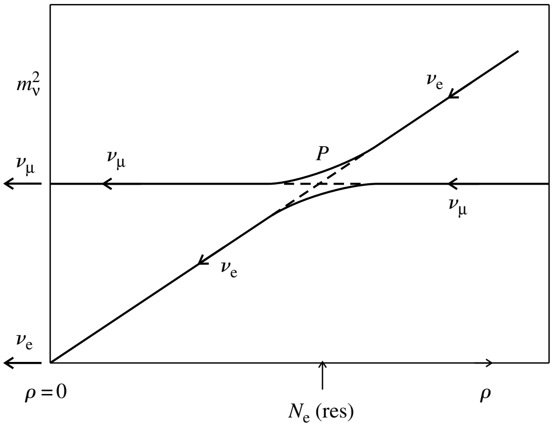


Fig. D.1 The MSW effect. The neutrino mass squared is plotted against solar density. For muon–neutrinos, the mass is independent of density and is represented by the horizontal line. For electron-neutrinos, the mass squared is proportional to the density as in (D.3) and, if there is no flavour mixing ($\theta = 0$) is represented by the diagonal line. The two levels cross at the point P, where $\Delta m^2 \cos 2\theta = 2\sqrt{2} G_{\mathrm{F}} N_{\mathrm{e}} E$ (see (D.7)). If the electron density in the solar core is greater than the 'resonance density' at P, the electron–neutrino will be located beyond P in the upper part of the diagram. As the neutrino moves outward into regions of lower density, it eventually reaches the resonance density, and provided the solar density varies slowly with radius, it will move along the continuous curve and emerge from the sun as a muon–neutrino.

described above, while the solar data on ν_e have been combined with those from $\bar{\nu}_e$ beams (assuming the validity of CPT), for example the KAMLAND experiment at the Kamioka mine site, using reactor antineutrinos on $\sim$200 km baseline. Of course, the reactor experiments measure the value of the vacuum mixing angle θ_{12} directly, and have short enough path lengths that they are untroubled by matter effects.

Answers to problems

Answers are given to all the problems. Fully worked solutions are given for those problems marked with an asterisk.

Chapter 1

(1.1) $\left(GM^2/R/2Mc^2\right) : 2.5 \times 10^{-14};\ 8 \times 10^{-21}$.

(1.2) In terms of quarks the reaction is written as follows:

$$\mathrm{d\bar{u} + udu \rightarrow uds + d\bar{s}}$$

The strong interactions have a range of order $r_0 = 1$ fm, hence a typical cross-section of $\sigma = \pi r_0^2 = 31$ mb ($= 3.1 \times 10^{-26}$ cm^2). The characteristic time is $r_0/c = 3 \times 10^{-24}$s. Thus a value of $\sigma = 1$ mb corresponds to a strong-interaction time in this case of 10^{-23} s. Hence the ratio of weak coupling to strong coupling is $\left(10^{-23}/10^{-10}\right)^{1/2} \sim 10^{-6}$.

(1.3) 29.8 MeV: 10 GeV: 5.7 GeV.

(1.4) 5.5×10^{-24} s; 134 fm.

(1.5) (a) yes; (b) and (c) no, $\Delta S = 2$ forbidden to first order in coupling; (d) no, because of energy conservation.

(1.6) If all final state lepton masses are neglected, the rate is proportional to Q^5. In nuclear beta decay, this is known as the *Sargent Rule*. The decay rate W in (1.15) has dimensions E^{-1}. The formula includes a factor E^{-4} from the weak coupling G_F^2 as in (1.27). Hence the other factors in the expression for W must have dimensions E^5, that is vary as Q^5 as Q is the important energy in the problem. The values of W/Q^5 in MeV^{-5} s^{-1} are as follows:

(a) 3.5×10^{-5} (b) 3.6×10^{-5} (c) 3.4×10^{-4} (d) 2.7×10^{-4} (e) 3.9×10^{-3}. The extreme relativistic approximation for the electron secondary does not hold for processes (c), (d), and (e) and W/Q^5 shows an increase with decreasing Q.

(1.7) We start by considering a massless neutrino of very high energy E, momentum p colliding with a nucleon of mass M at rest. The square of the energy in the CMS of the collision will be

$$s = (E + M)^2 - \left(p + 0\right)^2 = 2ME + M^2 \sim 2ME$$

If the quark carries a fraction x of the nucleon mass the result in the quark-neutrino CMS will be $s = 2xME$ and the cross-section from (1.27b) will be

$$\sigma = \frac{G_{\mathrm{F}}^2 s}{\pi} = \frac{2G_{\mathrm{F}}^2 xME}{\pi}$$

Inserting the values $G_F = 1.17 \times 10^{-5}$ GeV^{-2}, $M = 0.94$ GeV, 1 GeV$^{-1} = 0.1975 \times 10^{-13}$ cm (see Table 1.1) one obtains $\sigma = 3.2 \times 10^{-38}$ xE cm^2 where E is in GeV, or $\sigma = 0.8 \times 10^{-38} E$ cm^2 for $x = 0.25$. The actually measured high energy total neutrino cross-section per nucleon is $\sigma = 0.74E \times 10^{-38}$ cm^2.

(1.8) From (1.9) and (1.27) we can write for the differential cross-section

$$\frac{\mathrm{d}\sigma}{\mathrm{d}q^2} = \frac{g_w^4}{\left[\pi\left(q^2 + M_W^2\right)^2\right]}$$

where the four-momentum transfer squared has a maximum value $q^2(\max) = -s$, the square of the CMS energy. Hence the total cross-section, integrating from $q^2(\min) = 0$ to $q^2(\max)$ becomes

$$\begin{aligned}\sigma &= \left(\frac{g_w^4}{\pi}\right) \int \frac{\mathrm{d}q^2}{\left[-q^2 + M_W^2\right]^2} \\ &= \frac{g_w^4 s}{\left[\pi M_W^2\left(s + M_W^2\right)\right]} \to \frac{G_F^2 s}{\pi} \quad \text{for } s \ll M_W^2 \\ &\to \frac{G_{\mathrm{F}}^2 M_W^2}{\pi} \quad \text{for } s \gg M_W^2\end{aligned}$$

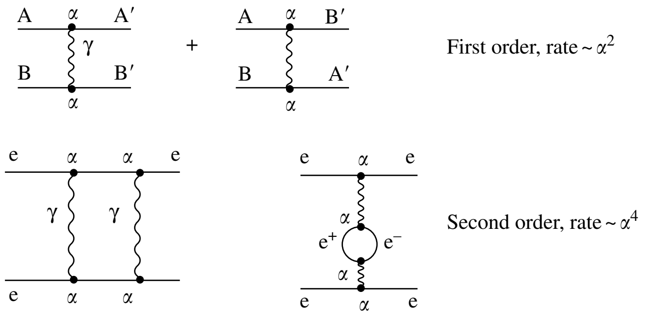
Inserting the values of the constants, the asymptotic cross-section equals 0.11 nb. The cross-section reaches half the asymptotic value when $s = M_W^2$, that is

$$E = \frac{M_W^2}{2m_e} = 6.3 \times 10^6 \mathrm{GeV}.$$

(1.10) 4×10^{-13} s.

(1.11) The three decays are identified with electromagnetic, weak and strong interactions respectively. If we set the strong coupling equal to unity, then from the data in the table, that for the electromagnetic interactions will be of order 1.6×10^{-2}, and that for weak interactions, of order 5×10^{-7}, taking the square roots of decay rates as proportional to the couplings.

(1.12) The diagrams are as follows:



In the first-order process of electron–electron scattering via single photon exchange, there are two diagrams depending how one labels the final-state particles as A or B. Since all that one observes is the scattered electron and not the vertices, both diagrams should be included.

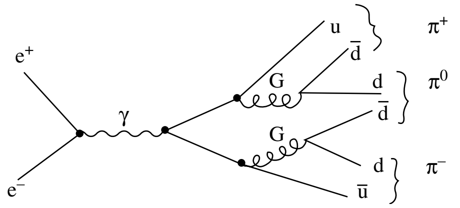
The second-order diagrams contain factors α^2 in amplitude or α^4 in rate, compared with α^2 for the first-order process, so they are relatively suppressed by a factor $\alpha^2 \sim 10^{-4}$.

(1.13) (a) and (b) are weak processes, (c) is electromagnetic and (d) is strong. Setting the strong coupling equal to unity, the weak and electromagnetic couplings are $\sim 10^{-8}$ and 10^{-3} respectively.

(1.14) The ratio $R = 3\sum (Q_i/e)^2$ where the factor 3 is for number of possible quark colours and the sum is over the charges Q_i of all relevant quark flavours. As a function of the CMS energy $\sqrt{s}$, the quark–antiquark flavours and values of R are as follows:

quarks	$\sqrt{s}$, GeV	R	
uū , dd̄	> 0.7	$3[(1/3)^2 + (2/3)^2]$	$= 5/3$
uū, dd̄, ss̄	> 1.0	$3[(1/3)^2 + (2/3)^2 + (1/3)^2]$	$= 6/3$
uū, dd̄, ss̄, cc̄	> 3.5	$3[(1/3)^2 + (2/3)^2$ $+(1/3)^2 + (2/3)^2]$	$= 10/3$
uū dd̄, ss̄, cc̄, bb̄	> 10	$3[(1/3)^2 + (2/3)^2 + (1/3)^2$ $+(2/3)^2 + (1/3)^2]$	$= 11/3$

The diagram for $e^+e^- \to \pi^+ + \pi^- + \pi^0$ is shown below. G represents a (strong) gluon exchange.



(1.15) The CMS energy of the pion in the decay $\Delta \to \pi + p$ is given by a little calculation in relativistic kinematics, as $E_\pi = \left(M_\Delta^2 + m_\pi^2 - M_p^2\right)/2M_\Delta = 0.267\text{GeV}$ (at the resonance peak). The corresponding pion momentum is $p_\pi = 0.228$ GeV/c. The CMS wavelength is then $\lambda_\pi = hc/p_\pi c = 8.6 \times 10^{-14}$cm. Inserting $J = 3/2$, $s_\pi = 0$, $s_p = 1/2$, $\Gamma_\gamma/\Gamma_{\text{total}} = 0.0055$, one finds $\sigma = 1.03$ mb. This is the cross-section for the reaction $\gamma + p \to \Delta$ at the resonance peak. In the head-on collision of a proton of high energy E_p with a photon of energy E_γ, the CMS energy squared will be $s = M_p^2 + 4E_\gamma E_p = M_\Delta^2$ if the collision excites the peak of the Δ resonance. The microwave radiation at $T = 2.73$ K has mean energy of 2.7 kT, and this corresponds to a quantum energy $E_\gamma = 6.3 \times 10^{-4}$ eV. Inserting in the above expression one obtains $E_p \sim 10^{21}$ eV. The mean free path of these protons through the microwave radiation will be $\lambda = 1/\rho\sigma$ where $\rho = 400\,\text{cm}^{-3}$ is the density of the microwave photons (see Chapter 5). Inserting the above value for the cross-section one obtains for the mean free path the value

$$\lambda = 2.5 \times 10^{22}\text{m} \sim 0.8 \text{ Mpc}.$$

(For further details, see Section 9.12)

Chapter 2

(2.1) See Section 2.8.
(2.2) 1.79×10^4 GeV^2
(2.4) Red shift, $\Delta\lambda/\lambda = +3.2 \times 10^{-4}$
(2.5) We refer to the transformations in Section 2.11. Assume the electron travels along the x-axis, and set $p_z = 0$ for convenience, so the transverse momentum is p_y. The angle of emission in the electron rest-frame is given by

$$\tan\theta^* = \frac{p_y^*}{p_x^*} = \frac{p_y}{\left[\gamma\left(p_x - \beta E/c\right)\right]}$$

where symbols with an asterisk refer to the electron rest-frame, and those without to the laboratory system, and $p_y^* = p_y$. With $p_y =$

$p \sin\theta$, $p_x = p\cos\theta$, and $E = pc$ for a photon one obtains

$$\tan\theta^* = \frac{\sin\theta}{[\gamma\,(\cos\theta - \beta)]}$$

In the electron rest-frame, half of the photons will have $\theta^* < \pi/2$, or $\cos\theta > \beta$, or $\sin\theta < 1/\gamma$. For ultra-relativistic particles, the half-width of the beam of emitted photons is therefore $\theta \sim 1/\gamma$.

(2.6) Gravitational shift $= +5.28 \times 10^{-10}$.
Special relativity shift $= -0.63 \times 10^{-10}$.
Net shift $= +4.45 \times 10^{-10}$.
$= 38\mu\text{s/day}$ (satellite clock runs fast).

Chapter 3

(3.2) Positive and negative pions are particle and antiparticle. Positive and negative sigma baryons are not.

(3.3) ρ- meson has $C = P = -1$. f-meson has $C = P = +1$. The process $\rho \to \pi^0\gamma$ is allowed as an electromagnetic decay, with a branching ratio $\sim \alpha$ (actually 0.07%). Corresponding decay for f-meson is forbidden by C-invariance.

(3.5) If $\mathbf{p_e}$, E_e, m_e, and $\mathbf{p_p}$, E_p,and M_p refer to the three-momenta, total energies and masses of the electron and proton respectively, then the square of the total four-momentum, equal to the CMS energy squared, is (see Section 2.11):

$$s = \left(E_\text{e} + E_\text{p}\right)^2 - \left(\mathbf{p_e} + \mathbf{p_p}\right)^2 = m_\text{e}^2 + m_\text{p}^2 + 2\left(E_\text{e}E_\text{p} - \mathbf{p_e}\cdot\mathbf{p_p}\right) \approx 4E_\text{e}E_\text{p}$$

where in the final step we have used the fact that both particles are extreme relativistic, so that masses can be neglected, and the fact that the electron and proton momenta are in opposite directions.

(a) Inserting numbers, the value of $s = 98,400\ \text{GeV}^2$.

(b) The CMS energy squared of the electron-quark system is $s/4$.

(c) The cross-section for the electromagnetic interaction is given by (1.23), which assumes that $q^2 \ll q^2_\text{max} = s$. In this approximation

$$\left(\frac{\mathrm{d}\sigma}{\mathrm{d}q^2}\right)_\text{em} = 4\pi\alpha^2\frac{|Q/e|^2}{q^4} \qquad \text{(i)}$$

where $|Q/e| = 2/3$ is the u-quark charge. The cross-section for the weak charged-current interaction is given by (1.23b) which, after allowing for the W propagator at high q^2 assumes the form

$$\left(\frac{\mathrm{d}\sigma}{\mathrm{d}q^2}\right)_{wk} = \frac{G_\mathrm{F}^2}{\left[\pi\left(1+\frac{q^2}{M_W^2}\right)^2\right]} \qquad \text{(ii)}$$

If we make the substitution $x = 3G_\mathrm{F}M_W^2/4\pi\alpha$ and set $\gamma = q^2/M_W^2$, then equating the above cross-sections gives the quadratic

$$\gamma^2\left(x^2-1\right) - 2\gamma - 1 = 0$$

with solution $\gamma = (1+x)/\left(x^2-1\right)$. Inserting numbers (see Appendix A) one obtains $x = 2.45$ and $\gamma = 0.69$, so that the cross-sections become equal at $q^2 = 4400\ \mathrm{GeV}^2$. Above this value, the charged weak current cross-section exceeds the electromagnetic cross-section.

[*Note:* the cross-section (i) has been stated in simplified form. At large q^2 it should be multiplied by a factor $\left[1+\left(1-q^2/q^2_\mathrm{max}\right)^2\right]/2$, but since the appropriate value of $q^2 \ll s$ this correction is small.]

(d) At large momentum transfers, neutral-current (Z exchange) as well as photon exchange in the process $e+p \to e+$ hadrons will become important.

(3.6) (a) Under interchange of space and spin coordinates, the wavefunction acquires a factor $(-1)^{L+S}$, that is $(-1)^S$ for a system with $L = 0$, and $S = 0$ or 1. But interchange of spatial and spin coordinates of electron and positron is equivalent to interchange of positive and negative charges, so that $C = (-1)^S$. If the positronium decays to two photons, it must have $C = +1$ so that this is the singlet state of $S = 0$, while decay to three photons implies $C = -1$ and $S = 1$. On account of the opposite parity of particle and antiparticle, the parity is $P = (-1)^{L+1} = -1$. Hence the quantum numbers are $J^\mathrm{PC} = 0^{-+}$ for the two-photon decay and 1^{--} for the three-photon decay.

(b) The energy levels are $E_n = -\alpha^2mc^2/\left(4n^2\right) = 6.806/n^2$ eV. The $n = 2 \to n = 1$ transition energy is $0.75 \times 6.806 = 5.1$ eV.

(c) The annihilation process needs the overlap of the electron and positron wavefunctions inside the volume they occupy, which is of the order of the cube of the Bohr radius $a = 2h/(mc\alpha)$. So for either decay, a factor $(m\alpha)^3$ enters the rate. The two-photon decay involves two lepton–photon vertices, hence a factor α^2, giving an overall factor $m^3\alpha^5$. A rate or width has dimensions of energy, hence dividing by m^2 to get the correct dimensions we can guess $\Gamma(2\gamma) \sim m\alpha^5$. In fact the true width is just half this, $m\alpha^5/2$. The three-photon decay clearly involves a third vertex and hence another factor of α. The full calculation yields $\Gamma\,(3\gamma) = \left[2\left(\pi^2-9\right)/9\pi\right]m\alpha^6$.

(3.7) $J^\mathrm{PC} = 1^{--}$. $\alpha_\mathrm{s} \sim 0.7$. (A more sophisticated analysis of upsilon levels gives $\alpha_\mathrm{s} \sim 0.2$).

(3.8) The transformations are as follows:

	T	P
$\mathbf{r}$	$\mathbf{r}$	$-\mathbf{r}$
$\mathbf{p}$	$-\mathbf{p}$	$-\mathbf{p}$
$\boldsymbol{\sigma}$	$-\boldsymbol{\sigma}$	$\boldsymbol{\sigma}$
$\mathbf{E}$	$\mathbf{E}$	$-\mathbf{E}$
$\mathbf{B}$	$-\mathbf{B}$	$\mathbf{B}$
$\boldsymbol{\sigma}\cdot\mathbf{E}$	$-\boldsymbol{\sigma}\cdot\mathbf{E}$	$-\boldsymbol{\sigma}\cdot\mathbf{E}$
$\boldsymbol{\sigma}\cdot\mathbf{B}$	$\boldsymbol{\sigma}\cdot\mathbf{B}$	$\boldsymbol{\sigma}\cdot\mathbf{B}$
$\boldsymbol{\sigma}\cdot\mathbf{p}$	$\boldsymbol{\sigma}\cdot\mathbf{p}$	$-\boldsymbol{\sigma}\cdot\mathbf{p}$

It is seen that an electric dipole moment for the neutron would violate both P and T invariance. So we can write for the dipole moment

$$\text{EDM} = \text{charge }(|e|) \times \text{length} \times P\text{-violating parameter}$$
$$\times\, T\text{-violating parameter}$$

The fact that P is violated means that we must introduce the weak coupling, with magnitude $G_{\mathrm{F}} = 1.17 \times 10^{-5}\,\mathrm{GeV}^{-2}$. We can get from this a characteristic length, which has dimensions 1/energy, by introducing a mass, which can be taken as the neutron mass. Thus with 1 GeV^{-1} = 1.97×10^{-14} cm, we find for the length $l = G_{\mathrm{F}}M_n \sim 2 \times 10^{-19}$ cm. For the T-violating parameter we assume the CPT theorem and take the equivalent CP violation rate from neutral kaon decay. The direct CP-violating rate is $\varepsilon' \sim 10^{-7}$, giving as our guess for the neutron electric dipole moment EDM $\sim 10^{-26}$ e cm. It is a pure accident that this is also the present (2007) experimental upper limit to the dipole moment. A full calculation with the Standard Model yields an estimate of 10^{-31} e cm, but other theories of 'physics beyond the Standard Model' yield values as high as 10^{-26} e cm.

A polarization asymmetry in proton-proton scattering, that is a dependence of the scattering cross-section on the sign of the beam helicity, would be a sign of parity violation. The expected level will be of the order of the ratio of weak-to-strong coupling amplitudes, that is of order 10^{-7}.

(3.9) All the decays are allowed, except for:

$$\rho^0 \to \pi^0 + \pi^0 \;\big(\text{forbidden by Bose symmetry, for which J must be even}\big)\,.$$

$$\rho^0 \to \pi^0 + \eta \;(C = -1 \to C = +1 \text{ transition forbidden in e.m. interaction})\,.$$

$$\eta \to e^+ + e^- \;(C = +1 \to C = -1 \text{ transition forbidden in e.m. interaction})\,.$$

The rate for $\pi^0 \to \gamma + \mathrm{e}^+ + \mathrm{e}^-$ is suppressed by a factor α relative to the two-photon decay.

(3.10) The pointlike cross-section (1.27b) is $\sigma = G_{\mathrm{F}}^2 s/\pi$ where G_F is the Fermi constant and s is the square of the centre-of-mass energy. The collision of a neutrino of energy E, negligible mass and momentum $pc = E$, with a stationary nucleon of mass M gives

$$\begin{aligned} s &= (E+M)^2 - (pc)^2 = M^2 + 2ME \\ &\approx 2ME \end{aligned}$$

for $E \gg M$ (units $h/2\pi = c = 1$). For collision with a parton with a fraction x of the nucleon four-momentum, the corresponding value of $s = 2MEx$. Hence, the average value of x will be

$$\langle x \rangle = \frac{\pi\sigma}{\left(2G_F^2 ME\right)}$$

In the above units, $G_F = 1.17 \times 10^{-5}$ GeV^{-2} (see Table 1.5), $M = 0.94$ GeV and σ is expressed in units of (0.1975 fm)^2—see Section 1.1. Inserting in the above expression yields $\langle x \rangle = 0.21$.

Chapter 4

(4.1) 30 mrad per year. 1.5×10^{20} years.
(4.2) 2.4×10^{32} year.
(4.3) $\Delta m^2 < 0.064$ eV^2.

Chapter 5

(5.1) Binding energy $\sim 10^{69}$ J. Mass energy $\sim 10^{70}$ J.
(5.2) $v^2 > 8\pi G\rho r^2/3$. Inserting $v = Hr$, the limit on the density is just the critical density (5.26)
(5.4) $(1 + z) = 107$. $T = 12$ million years, assuming matter domination for $z < 107$.
(5.5) 5750 K.
(5.6) $\varepsilon > 5 \times 10^{-19}$.
(5.7) For a radiation-dominated universe, $\rho = \left(3/32\pi G\right)/t^2$ (see (5.47)) while for matter domination $R = (6G\pi\rho R^3)^{1/3} t^{2/3}$(see (5.14)). After integration this gives for the time elapsed to reach a density ρ

$$t_{\text{rad}} = \left(\frac{3}{32\pi G\rho}\right)^{1/2}$$

$$t_{\text{mat}} = \left(\frac{1}{6\pi G\rho}\right)^{1/2}$$

which can be compared with the free-fall time of collapse of a body of density ρ from rest, (see Section 8.8):

$$t_{\text{freefall}} = \left(\frac{3\pi}{32G\rho}\right)^{1/2}$$

(5.8) Let the light signal start off at $t = t_1$, to reach us at $t = t_0$. Consider the time interval $\mathrm{d}t'$ where $t_1 < t' < t_0$. In this time interval the light signal covers a distance $c\mathrm{d}t'$, but by the time $t = t_0$, this will have expanded to $c\mathrm{d}t'R\,(0)/R\,(t')$ where $R(t')$ is the expansion factor at $t = t'$. We know that in a matter-dominated universe $R\,(0)/R\,(t') = (t_0/t')^{2/3}$. Hence the total distance travelled by the light signal will be:

$$L = R\,(0)\int \frac{c\mathrm{d}t'}{R\,(t')} = ct_0^{2/3}\int \frac{\mathrm{d}t'}{t'^{2/3}} = 3ct_0\left[1 - \left(\frac{t_1}{t_0}\right)^{1/3}\right]$$

The redshift is given by $(1 + z) = R\,(0)/R\,(t_1) = (t_0/t_1)^{2/3}$. Hence the time elapsed is

$$t_{\text{elapsed}} = 3t_0\left[1 - \frac{1}{(1 + z)^{1/2}}\right] = 0.85t_0.$$

(5.9) $< 2/3$, or $\Omega_\lambda > 1/3$.

(5.11) Referring to Equation (5.31) and Example 5.3, the expression for the age t_0 will be given by the integral

$$H_0t_0 = \int \frac{\mathrm{d}z}{(1 + z)\left[\Omega\,(1 + z)^3 + (1 - \Omega)\,(1 + z)^2\right]^{1/2}}$$

where the integral runs from $z = 0$ to $z = \infty$, and the $(1 + z)^3$ and $(1 + z)^2$ terms refer to matter and curvature contributions respectively, and $\Omega_k = 1 - \Omega$ where $\Omega \equiv \Omega_m$. To perform this integral, first make the substitution $(1 + z) = [(1 - \Omega)/\Omega]\tan^2\theta$, when it reduces to

$$A\left\{\int \mathrm{d}\theta\left[\frac{1}{\left(\sin^3\theta\right)} - \frac{1}{(\sin\theta)}\right]\right\}$$

$$= -\frac{1}{2}A\left\{\left(\frac{\cos\theta}{\sin^2\theta}\right) + \ln\left[\tan\left(\frac{\theta}{2}\right)\right]\right\}$$

where $A = 2\Omega/(1 - \Omega)^{3/2}$. The limits of integration are from $z = 0$, when $\tan^2\theta = \Omega/(1 - \Omega)$, to $z = \infty$, when $\tan^2\theta = \infty$ and $\theta = \pi/2$. Inserting the numerical value $\Omega = 0.24$, one obtains $H_0t_0 = 0.832$ or $t_0 = 11.3$ Gyr.

(5.12) The stages in the answer are as follows:

(1) Inserting the value of the Fermi constant, and with s in MeV^2, the cross-section is found to be $\sigma = 2.82 \times 10^{-45}$ s cm^2.

(2) If the momentum and energy of electron and positron are $\mathbf{p}_1 (= E_1)$ and $\mathbf{p}_2 (= E_2)$, then $s = (E_1 + E_2)^2 - (\mathbf{p}_1 + \mathbf{p}_2)^2 = 2E_1E_2(1 - \cos\theta)$ where θ is the angle between the two momenta. These are isotropically distributed, so that $<\cos\theta> = 0$ and $<s> = 2<E>^2 = 2 \times (3.15kT)^2$.

(3) Referring to (5.56), the number density of electrons or positrons is $N_\mathrm{e} = (3/4) \times 2.404 \times (kT)^3/\left(\pi^2h^3c^3\right) = 2.39 \times 10^{31}\,(kT)^3\ \mathrm{cm}^{-3}$

(with kT in MeV). Setting the relative velocity $v \sim c$, the value of $1/W = 1/\langle \sigma N_e v \rangle \sim 25/(kT)^5$s.

(4) From (5.49) and (5.58) one obtains t=$0.74/(kT)^2$s for g^*=43/4. Equating t with $1/W$ gives for the freeze-out temperature kT~3 MeV.

[*Note*: This is only an approximate value, for several reasons. First, the calculation of both cross-sections and particle densities assumes all particles are extreme relativistic, and the corrections for electron mass will reduce the cross-section, the particle density and the relative velocity and thus increase the freeze-out temperature. Second, the calculation of the cross-section ignores the effects of Z^0-exchange (neutral weak currents) which will increase it by about 15% and hence reduce the critical temperature].

Chapter 6

(6.1) As indicated in Chapter 5, the freeze out of neutrons and protons from equilibrium occurs when the interaction rate W in (6.1), varying as T^5, falls below the expansion rate H in (5.59), varying as T^2 $g^{*^{1/2}}$. Thus the freeze-out temperature $T \propto g^{*^{1/6}}$, where $g^* = (22 + 7N_\nu)/4$ is the number of states of photons, electrons, positrons, and neutrinos/antineutrinos and N_ν is the number of neutrino families (see Section 5.10). For $N_\nu = 3$, $g^* = 43/4$, $kT = 0.8$ MeV, so that kT for other values of N_ν is easily found. Inserting in (6.4), the initial and final neutron/proton ratios and hence the helium mass fraction can be calculated as a function of the assumed number of neutrino families.
For a neutron–proton mass difference of 1.4 MeV and three neutrino flavours, the initial neutron/proton density ratio in (6.5) becomes 0.174, leading to a helium mass fraction of 0.21.

(6.2) Referring to equation (1.18) the cross-section for the reaction $\nu_e + n \to e^- + p$ is given by the approximate formula (putting $|T_{if}|^2 = G_F^2$):

$$\sigma = \left(\frac{1}{\pi \hbar^4 c^4}\right) G_F^2 \left(p_f c\right)^2$$

where we have taken $v_i = v_f = c$. With $G_F = 1.17 \times 10^{-5}\, (\hbar c)^3\, \text{GeV}^{-2}$ this gives $\sigma = 1.69(p_f c)^2 \times 10^{-44}$ cm^2 with the final state momentum $p_f c$ in MeV. The width or rate for the reaction per neutron target is found from (1.14) as $W = \sigma\phi$ where $\phi = nc$ is the flux of incident neutrinos and n is their number density as given by (5.50) and (5.56), that is $\phi = 2 \times 10^{42} (kT)^3$ in units cm^{-2} s^{-1}, with kT measured in MeV. This gives a value of $W = 0.05(kT)^3 (p_f c)^2$ s^{-1}, to be compared with $H = 0.7$ (kT)^2 s^{-1} from (5.59) for three neutrino flavours. The value of the final state momentum in the reaction will be $p_f c \sim kT + Q$ where $Q = 1.29$ MeV. Setting $W/H = 1$ for the freeze-out condition gives $kT \sim 1.5$ MeV, as can be found by trial and error. This is an overestimate of the freeze-out temperature for several reasons. First, in assuming equation (1.18) we have ignored the effects of spin, and that both vector and axial vector interactions are involved in the matrix element. This will increase the cross-section and decrease the freeze out temperature; second, we have ignored the electron mass, which with a value of 0.51 MeV is comparable with kT; third, an integration should be made over the thermal spectrum and over the isotropic angular distribution of the colliding particles. When all such effects are accounted for, the freeze-out temperature is found to be approximately 0.8 MeV.

(6.3) $kT = 0.06$ MeV

(6.4) 5%

Chapter 7

(7.1) $\Delta\theta = 2\mathrm{G}M/bc^2$

[*Note*: This is too small by a factor 2, as compared with the value given by the general theory of relativity. One reason for the discrepancy is that, in the Newtonian approach, only the spatial coordinates of the photon are considered, while in fact the gravitational field also affects the time coordinate, as explained in Chapter 2. This introduces a time delay (called the Shapiro delay) which must introduce an extra deflection (as is familiar in classical optics, when the speed of light changes as it travels from a less dense to a more dense medium).]

(7.2) $v^2 = \dfrac{\mathrm{G}M}{R}$

$v = 210\text{kms}^{-1}$

Optical depth $\tau = v^2/c^2 = 10^{-6}$.

(7.3) $E_\mathrm{R} = \left[\dfrac{4M_\mathrm{D}M_\mathrm{R}}{(M_\mathrm{D}+M_\mathrm{R})^2}\right] E_\mathrm{D}\cos^2\theta$

E_R is maximum when $\theta = 0$.

$E_\mathrm{R}(\max) = E_\mathrm{D}$ when $M_\mathrm{R} = M_\mathrm{D}$.

For $M_\mathrm{R} \ll M_\mathrm{D}$, $E_\mathrm{R}(max) = 4E_\mathrm{D}\left(\dfrac{M_\mathrm{R}}{M_\mathrm{D}}\right)$

For $M_\mathrm{D} \ll M_\mathrm{R}$, $E_\mathrm{R}(max) = 4E_\mathrm{D}\left(\dfrac{M_\mathrm{D}}{M_\mathrm{R}}\right)$.

$E_\mathrm{R} \sim 70$ keV in numerical problem.

(7.4) $1.03 \times 10^{-11}\ \mathrm{m\ s^{-2}} \sim 10^{-12}$ g.

Chapter 8

(8.1) $\tau = \dfrac{2\pi}{\omega} = \left(\dfrac{3\pi}{\mathrm{G}\rho}\right)^{1/2}$

(8.2) $v_s = 360\ \mathrm{m\ s^{-1}}$,

$\lambda_J = 6.88 \times 10^7$ m,

$M_\mathrm{J} = \dfrac{\pi\rho\lambda_\mathrm{J}^3}{6} = 2.2 \times 10^{23}$kg

(cf Earth mass $= 6 \times 10^{24}$kg)

(8.4) $t = 3\lambda^2/(c^2 t_i)$

$$\lambda = 1\text{cm} \quad t = 10^4\text{y} \quad M = 10^{16} M_S \sim \text{cluster mass}$$

$$\lambda = 1\text{mm} \quad t = 10^2\text{y} \quad M = 10^{12} M_S \sim \text{galaxy mass.}$$

(8.6) 14.9 Gpc.

(8.7) Consider photons travelling through ionized gas in a flat, expanding universe. From (5.43) and (5.44) the element of (true) coordinate distance travelled is related to the red shift interval dz by

$$\mathrm{d}D = \left(\frac{c}{H_0}\right) \frac{\mathrm{d}z}{\begin{gathered}\left[\Omega_m(0)(1+z)^3 + \Omega_r(0)(1+z)^4\right. \\ \left. + \Omega_v(0) + \Omega_k(0)(1+z)^2\right]^{1/2}\end{gathered}}$$

Since we will be concerned with z values of around 20 or less, we can to a first approximation neglect all terms in the denominator of the expression for dD except the first. Then the probability that a photon will undergo Thomson scattering in traversing this distance will be

$$\mathrm{d}P = \left[\Omega_b(0)(1+z)^3 \rho_c\right] \sigma_T N_0 \mu(z)\, \mathrm{d}D$$

where the product in square brackets gives the baryon—and hence electron—density at redshift z. ρ_c is the critical density today, σ_T is the Thomson cross-section for photon–electron scattering in the eV energy region, N_0 is Avogadro's number, and $\mu(z)$ is the degree of ionization of the intergalactic medium. The integrated scattering probability from $z = z$ to $z = 0$ becomes

$$P = \left(\frac{c}{H_0}\right) \frac{\Omega_b(0)}{[\Omega_m(0)]^{1/2}} \rho_c \sigma_T N_0 \int (1+z)^{3/2} \mu(z)\, \mathrm{d}z$$

We do not know $\mu(z)$, except that it certainly decreases as z increases. For simplicity, let us assume that stars formed at an effective value of z = z_s, below which the gas is 100% ionized, so that for $z < z_s$, $\mu(z) = 1$, and $\mu(z) = 0$ for $z > z_s$. Then, assuming $P = 0.1$ (the value quoted by the WMAP experiment) and inserting the values of the other constants, one finds $z_s = 12$, which translates to a time of ~ 0.25 Gyr after the Big Bang.

(8.8) $v_{\text{galaxy}} \sim 2.5 \times 10^5$ m s^{-1}; $v_{\text{cluster}} \sim 4 \times 10^5$ m s^{-1}
Neutrino r.m.s. velocities 2.1×10^7, 6.8×10^6, and 2.1×10^6 m s^{-1} respectively.

(8.9) From (5.32) we take for the present baryon density, assuming it to be dominated by protons, the value of 0.24 protons m^{-3}, which is also the electron density $N_e(0)$. At the time of decoupling ($z_{\text{dec}} \sim 1100$) the electron density would be $N_e(\text{dec}) = N_e(0)(1 + z_{\text{dec}})^3$. The Thomson cross-section (1.26d) is $\sigma_{\text{Th}} = 6.7 \times 10^{-29}$ m^2, giving a mean free path for CMB photons of $\lambda_{\text{Th}} = 1/(N_e \sigma) = 4.7 \times 10^{19}$m $= 1.51$kpc, assuming the medium to be totally ionized. Since near $z = 1100$, baryonic matter is only partly ionized, we take a value of 10λ as illustrative of the thickness of this 'last scattering shell'.

From (5.43) and (5.44), and assuming a flat universe ($k = 0$) with dominance of matter over radiation and vacuum energy, as expected for $z \sim 1100$, we obtain the value of the shell thickness ΔD related to the corresponding interval Δz as follows:

$$\Delta D \sim \left(\frac{c}{H_0}\right) \mathrm{d}z \left[\Omega_m(0)\,(1 + z_{\mathrm{dec}})^3\right]^{-1/2} = 8.1 \times 10^{21} \Delta z \,\mathrm{m}$$

Equating this to the present thickness of the last scattering shell, 10 $\lambda_{\mathrm{Th}}(1 + z_{\mathrm{dec}}) = 5.16 \times 10^{23}$ m, one obtains $\Delta z \sim 64$. This corresponds to an angular uncertainty $\Delta\theta \sim 0.03°$, from (8.64), and the smearing out of peaks at such small angles.

Chapter 9

(9.1) $E^{-2.22}$.

(9.2) 110 m.

(9.3) 2.4 km.

(9.4) The probability that a pion is produced in an element dx at depth x g cm^{-2} in the atmosphere by a primary proton is $\exp(-x/\lambda)\,\mathrm{d}x/\lambda$, where the interaction length $\lambda \sim 100$ g cm^{-2}. The probability that it will then survive to depth y without interaction is $\exp\left[-(x - y)/\lambda\right]$, where to keep the problem simple we assume the same interaction length for pions and protons.

The connection between the depth x in g cm^{-2} and height h is $h = H \ln(X/x)$, where $X = 1030$ g cm^{-2} is the total atmospheric depth, and again for simplicity we have assumed an isothermal, exponential atmosphere (strictly only true in the top third of the atmosphere) with $H = 6.5$ km.

In traversing the interval x to y the pion covers a distance $s = H \ln\left(y/x\right)$, and the probability that it does this without decaying, and subsequently decays in an element ds is

$$\mathrm{d}P = \exp\left(\frac{-s}{\gamma c\tau}\right)\frac{\mathrm{d}s}{\gamma c\tau} = \left(\frac{H}{\gamma c\tau}\right)\cdot\left(\frac{\mathrm{d}y}{y}\right)$$
$$\cdot \exp\left[-\left(\frac{H}{\gamma c\tau}\right)\ln\left(\frac{y}{x}\right)\right] = \alpha\left(\frac{x}{y}\right)^{\alpha}\frac{\mathrm{d}y}{y}$$

where $\alpha = H/\gamma c\tau = Eo/E$. Here E is the pion energy, $\gamma = E/mc^2$ and m and τ are the pion mass and mean lifetime. Thus the joint probability that a pion is created in dx and survives to decay in the depth interval dy is

$$P\left(x, y\right)\mathrm{d}x\,\mathrm{d}y = \left(\frac{\mathrm{d}x}{\lambda}\right)\exp\left(\frac{-x}{\lambda}\right)\exp\left[-\frac{\left(x - y\right)}{\lambda}\right]\alpha\left(\frac{x}{y}\right)^{\alpha}\frac{\mathrm{d}y}{y}$$

and the probability that the pion is produced at *any* value of $x < y$ and decays in dy is found by straightforward integration to be

$$P\left(y\right)\mathrm{d}y = \left[\frac{\alpha}{(\alpha + 1)}\right]\exp\left(-\frac{y}{\lambda}\right)\frac{\mathrm{d}y}{\lambda}$$

The overall probability that the pion decays anywhere in the atmosphere is found by integrating from $y = 0$ to $y = X$. Since $X \gg \lambda$ the y integration just gives unity. So the overall pion decay probability is

$$P_{\text{decay}} = \frac{\alpha}{(\alpha + 1)} = \frac{E_0}{(E + E_0)}$$

where $E_0 = Hmc^2/c\tau = 117\text{GeV}$.

If the pion is produced at a zenith angle θ, depths remain the same, but all distances are multiplied by the secant of this angle, so that the energy E_0 is simply replaced by $E_0\sec\theta$.

(9.5) No CP-violating effects are possible with only two flavours in vacuum, since a CP-violating phase requires at least three flavours and a 3 ×3 mixing matrix. If matter effects in the Earth are taken into account, induced CP-violating effects, that is a difference in the oscillation amplitudes for neutrinos and antineutrinos are possible with only two neutrino flavours, because the Earth is not CP-symmetric, being made of matter without antimatter. (see also Appendix D).

(9.6) Applying the conservation of energy and the conservation of momentum, parallel, and perpendicular to the direction of the incident neutrino, allows one to eliminate the energy and angle of emission of the recoil electron, and there results a relation between the angle of emission θ of the scattered neutrino and its energy, E', in terms of the incident energy E:

$$\cos\theta = 1 - m\left(\frac{1}{E'} - \frac{1}{E}\right)$$

where m is the electron mass, and the neutrino is assumed massless. Since $E \gg m$, we can expand $\cos\theta \approx 1 - \theta^2/2$, whence we obtain the relation

$$\theta = \sqrt{2m\left(\frac{1}{E'} - \frac{1}{E}\right)}$$

(9.7) Applying the energy and momentum transformations in Section 2.11, one obtains values for the energy and momentum, E^* and p^*, of the muon in the pion rest-frame:

$$E^* = \frac{\left(m_\pi^2 + m_\mu^2\right)}{2m_\pi}; \quad p^* = \frac{\left(m_\pi^2 - m_\mu^2\right)}{2m_\pi}$$

and for the laboratory energy of the muon from decay of a relativistic pion of Lorentz factor $\gamma = E_\pi/m_\pi$ and $\beta \approx 1$ the value

$$E_\mu = \gamma\left(E^* + p^* \cos\theta^*\right)$$

where θ^* is the angle of emission of the muon in the pion rest-frame. Because the pion has spin zero, this angular distribution is isotropic, and the muon energy in the laboratory therefore extends from $\left(m_\mu^2/m_\pi^2\right)E_\pi = 0.58E_\pi$ to E_π, with a mean value of $0.79E_\pi$. Hence the neutrino receives an average energy of $0.21E_\pi$. In its subsequent decay, the muon transforms into a positron, an electron–neutrino and a muon–antineutrino, each receiving approximately one-third of the muon energy, that is about $0.26E_\pi$. In summary, therefore, the average energies of the various neutrinos are as follows:

$$\pi^+ \to \mu^+ + \nu_\mu \qquad \langle E(\nu_\mu)\rangle = 0.21E_\pi$$
$$\mu^+ \to \mathrm{e}^+ + \nu_e + \bar{\nu}_\mu \qquad \langle E(\nu_e)\rangle = \langle E(\bar{\nu}_\mu)\rangle = 0.26E_\pi$$

The numbers here ignore the effects of spin polarization of the muons from pion decay, which can affect the mean energies by several percent.

(9.8) $P = 5$ microwatts.

(9.10) 3×10^{-4}.

(9.13) Suppose the jet of relativistic particles emits a light signal at time t_0 and a second signal at time $(t_0 + \Delta t)$. Taking the x-axis as the line of sight to the Earth and the y-axis in the transverse direction, the actual transverse velocity of the jet is $\Delta y/\Delta t = v \sin\theta$, but this is not the value observed at the Earth. Since the jet is moving towards the Earth with velocity $v\cos\theta$, the time on the Earth between the two signals is

$$\Delta t_{\mathrm{E}} = \Delta t - \frac{\Delta x}{c} = \Delta t\left[1 - \left(\frac{v}{c}\right)\cos\theta\right].$$

Hence the apparent transverse velocity measured on the Earth is

$$\frac{u_{\mathrm{trans}}}{c} = \frac{\Delta y}{c\Delta t_{\mathrm{E}}} = \frac{\beta\sin\theta}{(1-\beta\cos\theta)}$$

where $\beta = v/c$ and $\gamma = 1/\sqrt{(1-\beta^2)}$. Differentiation shows that u_{trans}/c has a maximum value of $\gamma\beta$ when $\sin\theta = 1/\gamma\beta$, and therefore exceeds unity when $\beta > 1/\sqrt{2}$. On the contrary, when $\theta > \pi/2$, the 'away jet' will be observed to have a transverse velocity less than β sinθ.

Chapter 10

(10.1) $\omega = 0.63$ rd s^{-1}.

(10.2) 5 billion years.

(10.3) If the mass M is to be supported by degeneracy pressure, the density is given by (10.29):

$$\rho_{\mathrm{deg}} = \left(\frac{4m_e^3}{h^6}\right)\left(\frac{Am_{\mathrm{P}}}{Z}\right)^5\left(\frac{4\pi}{3}\right)^3 M^2G^3$$

If M is small enough, ρ_{deg} will fall below normal solid matter densities and atomic (electromagnetic) forces will then prevent gravitational collapse. So the maximum mass not dependent on electron degeneracy for stability is found by setting $\rho_{\mathrm{deg}} = \rho_{\mathrm{matter}} = 10^4$ kg m^{-3}. Inserting the various constants yields $M \sim 5 \times 10^{27}$ kg or about 0.25 % of the solar mass. The largest planet in the solar system is Jupiter with $M = 0.001M_{\mathrm{sun}}$, for which the increase in central density due to electron degeneracy would be only about 10%.

(10.4) $t < 660$ years, inconsistent with its origin in 1054 A.D.

(10.5) The particle horizon distance is nct_0 where $n = 2$ for a radiation-dominated universe and $n = 3$ for matter-domination. If the mass is M, then

$$r_s = \frac{2GM}{c^2} = nct_0$$

$$\text{or} \quad M_{BH} = \frac{nc^3t_0}{2G}$$

The mass of the universe of age t_0 and critical density is $(4\pi/3)\,\rho_c\,(nct_0)^3$. Inserting the values in SI units of $\rho_c = 9 \times 10^{-27}$ and $G = 6.7 \times 10^{-11}$, this is equal to the above mass when $t_0 = 1.5 \times 10^{10}/n$years.

(10.6) $R \sim 10^{-15}$ m (about equal to the radius of a proton).
$M \sim 10^{12}$ kg (about equal to the mass of a typical mountain).
(Inserting all the constants in (10.51) gives $t = 8.1 \times 10^{66}$ $(M/M_{\text{sun}})^3$ years. And equating to the age of 14 billion years gives $M = 2.41 \times 10^{11}$ kg and $R = 7.4 \times 10^{-16}$ m.)

(10.8) For an elliptic orbit with semi-major axis a and eccentricity e, the velocity of the star in orbit at radius vector r is given by $v^2 = GM\,[(2/r) - 1/a]$ where M is the mass of the black hole, situated at the focus of the ellipse. At the perigee, $r = a(1 - e)$ and $v^2 = (GM/a)\,[(1+e)/(1-e)]$, while the period of the orbit is given by Kepler's Law as $\tau^2 = 4\pi^2a^3/[GM]$. Inserting the value of the eccentricity $e = 0.87$, period $\tau = 15 \textit{years} = 4.7 \times 10^8$ s, and the perigee distance as 17 lighthours $= 1.84 \times 10^{13}$ m, we obtain $a = 1.4 \times 10^{14}$ m, $M = 3.65 \times 10^6 M_{\text{sun}}$, and the velocity of the star at perigee as $v = 7170$ km per s.

This page intentionally left blank

References

Abassi R.U. *et al. Astroph. J.* **622**, 910 (2005)

Abassi R.U. *et al. astro-ph/0703099* (2007)

Abbott B. *et al. Phys. Rev.* **D73**, 062001 (2006)

Abraham J. *et al. Nucl. Instr. Methods* **A523** 50 (2004)

Abraham J. *et al.* Pierre Auger Collaboration, *Science* **318**, 938 (2007)

Adler R., M. Bazin, and M. Schiffer *"Introduction to General Relativity"* (McGraw-Hill 1965)

Akerib D.S. *et al. Phys. Rev.* **D73**, 011102(2006); *Phys. Rev. Lett.* **96**, 011302 (2006)

Albrecht A. and P.J. Steinhardt, *Phys. Rev. Lett*. **48**, 1220 (1982)

Alcock C. *et al. Nature* **365**, 621 (1993)

Alekseev E.N. *et al. JETP Lett*. **45**, 589 (1987)

Allen D. *et al. Monthly Notices of the RAS* (18/5/2004)

Alvarez M.A. *et al. Astroph. J.* **644**, L101 (2006)

Anderson C.D. *Phys. Rev*. **43**, 491 (1933)

Anderson H.L. *et al. Phys. Rev*. **85**, 934 (1952)

Arnison G. *et al. Phys. Lett*. **122B**, 103 (1983)

Bahcall J.N. *"Neutrino Astrophysics"* (Cambridge University Press 1989)

Bahcall J.N., H.M. Pinsonneault, and S. Basu *Astroph. J.* **555**, 990 (2001)

Barnett R.M. *et al.* (*Review of Particle Physics*) *Phys. Rev*. **D54**, 1 (1996)

Barrow J.D. *Quart. J. Roy. Astr. Soc*. **29**, 101 (1988)

Bathow G. *et al. Nucl. Phys*. **B20**, 592 (1970)

Bennett C.L. *et al. astro-ph*/0302207 (2003)

Bennett C.L. *et al. Astroph. J.* **464**, L1 (1996)

Benoit A. *et al. astro-ph* /0206271 (2002)

Bernabei R. *et al. Phys. Lett*. B480, 23 (2002)

Bernabei R. *et al. arXiv:0804.2741* (2008)

Bionta R.M. *et al. Phys. Rev. Lett.* **58**, 1494 (1987)

Bondi, H. and R. Littleton, *Nature* **184**, 974 (1959)

Bosetti P.C. *et al. Nucl. Phys.* **B142**, 1(1978); **B203**, 362(1982)

Braginsky V.B. and V.I. Panov, *Sov. Phys. JETP* 34, 463 (1972)

Broeils A.H. *Astr. and Astroph*. **256**, 19 (1992)

Buks E. and M.L. Roukes *Nature* **419**, 119 (2002)

Cabbibo N. *Phys. Rev. Lett*. **10**, 531 (1963)

Cameron R. *et al.*, *Phys Rev* **D47**, 3707 (1993)

Casimir H.B.G. *Proc. Kon. Ned. Akad*. **51**, 793 (1948)

Chaboyer B. *Nucl. Phys*. **B51**, 11 (1996)

Chandrasekhar S. *Astroph. J.* **74**, 81 (1931)

Christenson J.H. *et al. Phys. Rev. Lett*. **13**, 138 (1964)
Clochiatti A. *et al. Astroph. J*. **642**, 1 (2006)
Clowe D. *et al. Astroph. J*. **648**, L109, (2006)
Cranshaw T. and A. Hillas, *Nature* **184**, 892 (1959)
Cronin J.W. *Rev. Mod. Phys*. **S166** (1999)
de Bernardis P. *et al. Astroph. J*. **564**, 539 (2002)
de Lapparent *et al. Astroph. J. Lett*. **302**, L1 (1986)
Dermer C.D., *Proc. 30th Int.Cosmic Ray Conf.*, Merida, Mexico (2007)
Dirac P.A.M., *Proc Roy Soc* **A133**, 60 (1931)
Doroshkevich A. *et al. astro-ph*/0307233 (2003)
Davis R., *Phys. Rev. Lett*. **12**, 303 (1964)
Davis R., *Prog. Part. Nucl. Phys*. **32**, 13 (1994)
Ehret K. *et al. arXiv:hep-ex*/0702023 (2007)
Elgaroy O. *et al. Phys. Rev. Lett*. **80**, 061301 (2002)
Enge H.A. *"Introduction to Nuclear Physics"* (Addison-Wesley 1972)
Feynman R.P., *Phys. Rev. Lett.* **23**, 1415 (1969)
Fiorini E. *Physica Scripta* **T121**, 86 (2005)
Fixen D.J. *et al. Astroph. J*. **473**, 576 (1996)
Friedmann H.A . *Zeit. Physik* **10**, 377 (1922)
Freedman J.L. *et al. Astroph. J*. **553**, 47 (2001)
Friedman J.T. and H.W. Kendall *Ann. Rev. Nucl. Part. Sci*. **22**, 203 (1972)
Fukuda Y. *et al. Phys. Lett*. **436**, 33 (1998); **433**, 9 (1998)
Fukugita M. and Yanagida T., *Phys. Lett.* **B174,** 45 (1986)
Gell-Mann M. *Phys. Lett*. **8**, 214 (1964)
Georgi H. and S.L. Glashow *Phys. Rev. Lett*. **32**, 438 (1974)
Glashow S.L. *Nucl. Phys*. **22**, 579 (1961)
Gribov V. and B. Pontecorvo, *Phys. Lett.* **B28**, 493 (1969)
Greisen K. *Phys. Rev*. **118**, 316 (1966)
Gundlach, J.H. *et al. Phys. Rev. Lett*. **78**, 2523 (1997)
Guth A.H. *Phys. Rev*. **D23**, 347 (1981); *Phys. Rep*. **333**, 555 (2000)
Halverson N.W. *et al. astro-ph*/0104489 (2001)
Han J.L. *et al. Astroph. J*. **642**, 868 (2006)
Hawking S.W. *Nature* **248**, 30 (1974)
Hewish A. *et al. Nature* **217**, 709 (1968)
Higgs P.W. *Phys. Lett*. **12**, 132 (1964); *Phys. Rev*. 145, 1156 (1966)
Hirata K.S. *et al. Phys. Rev. Lett.* **58**, 1490 (1987)
Hoyle F. *Astroph. J. Suppl*. **1**, 121 (1954)
Hulse R.A. and J.H. Taylor *Astroph. J*. **191**, L59 (1974): **201**, L55 (1975)
Itzykson C. and J.B. Zuber *"Quantum Field Theory*" (McGraw-Hill 1985)
Kamionkowski M. and A. Kosowski *Ann. Rev. Nucl. Part. Sci.* **49**, 77 (1999)
Kapner D.J *et al. Phys. Rev. Lett*. **98**, 021101(2007)
Klapdor-Kleingrothaus H.V. *et al. J. Phys. J*. **12A**, 147 (2001)
Kneib J.P. *et al. Astroph. J*. **471**, 643 (1996)
Kobayashi M. and K. Maskawa *Prog. Theor. Phys*. **49**, 282 (1972)
Koks F.W.J. and J. Van Klinken *Nucl. Phys*. **A272**, 61 (1976)

Kolb E.W. *Proc. 29th Int. Conf. High En. Phys.* (Vancouver 1998, ed. A. Astbury *et al.*)

Kolb E.W. and M.S. Turner "*The Early Universe"* (Addison-Wesley 1990)

Kronberg P.P. *J. Korean Astron. Soc*. **37**, 343 (2004)

Kuzmin V.A. *et al. Phys. Lett.* **155**, 36 (1985)

Lamoreaux S.K. *Phys. Rev. Lett*. **78**, 5 (1997)

Lee A.T. *et al. astro-ph*/0104459 (2001)

Lehraus I. *et al. Nucl. Instr. Meth*. **153**, 347 (1978)

Linde A.D. *Phys. Rev. Lett*. **B108**, 389 (1982); *Rep. Prog. Phys*. **47**, 925 (1984)

Lyne A.G. *et al. Science* **303**, 1153 (2004)

Maki Z., M. Nakagawa, and S. Sakata, *Prog. Theor. Phys*. **28**, 870 (1962)

Mather J.C. *et al. Astroph. J*. **354**, L37(2000)

Mikhaev S.P. and A.Y. Smirnov *Nuov. Cim*. **9C**, 17 (1986)

Ong R.A. *Phys. Rep*. 305, 93 (1998)

Peacock J.A. "*Cosmological Physics*" (Cambridge University Press 1999)

Peccei R. and H. Quinn *Phys. Rev. Lett*. **38**, 1440 (1977)

Penzias A.A. and R.W. Wilson *Astroph. J.* **142**, 419 (1965)

Perkins D.H. *Proc. XVI Intl. Conf. High Energy Phys*. Vol. 4, 189 (Fermilab 1972)

Perlmutter S. *et al. Astroph. J*. **517**, 565 (1999)

Politzer H.D. *Phys. Rep*. **14C**, 129 (1974)

Pontecorvo B., *J. Exp. Theor. Phys*. **53**, 1717 (1967)

Pound R.V. and J.L. Snider, *Phys. Rev. Lett*. 13, 539 (1964)

Ressell M.T and M.S. Turner, *Comments in Astrophysics* **14**, 323 (1990)

Riess A.G. *et al. Astron. J*. **116**, 1009 (1998)

Riess A.G. *et al. Astroph. J*. **536**, 62 (2000)

Riess A.G. *et al. Astroph. J*. **607**, 665 (2004)

Rolfs C.E. and W.S. Rodney *"Cauldrons in the Cosmos*" (University of Chicago Press, Chicago 1988)

Roy A. *et al. Phys. Rev*. **D60**, 111101 (1999)

Sakharov A. *JETP Lett.* **5**, 241 (1967)

Salam A. *Elementary Particle Theory* (Stockholm: Almquist and Wiksell 1967)

Schramm D.N. and M.S. Turner *Rev. Mod. Phys*. **70**, 303 (1998)

Simpson J.A. *Ann. Rev. Nucl. Part. Sci*. **33**, 326 (1983)

Smith N.J.T. (*UK DMC report at Dark Matter Conference*, York July 2002)

Smoot G. F. *et al. Astroph. J*. **360**, 685 (1990)

Smy M. *(Super-Kamiokande Collaboration) APS meeting*, UCLA, Jan. 1999

Sparnaay M.J. *Physica* (Utrecht) **24**, 751 (1958)

Spergel D.N. *et al. Astroph. J. Suppl Series* **148**, 175 (2003)

Stoner E.C., *Phil. Mag*. **7**, 63 (1929)

Stoner E.C., *Phil. Mag*. **9**, 944 (1930)

Suntzeff N.B. *et al. Astroph. J. Lett*. **384**, L33 (1992)

Surdej J. *et al. Nature* **329**, 695 1987)

Suzuki Y., *Physica Scripta*, **T121**, 23 (2005)

Tannenbaum M.J., *Rep. Prog. Phys*., **69**, 2005 (2006)

Tegmark M. *et al. Astroph. J*. **606**, 702 (2004)

Tegmark M.*Phyica. Scripta* **T121**, 153 (2005)

‘t Hooft G., *Phys. Rev. Lett.* **37**, 8 (1976)

‘t Hooft G., *Phys. Rev*. **D14**, 3432 (1976)

Wagner R. *et al. Proc. 29th Int. Conf. Cosmic Rays*, Pune, Aug. 2005, Vol.4 p.163

Webber W.R. *Nuov. Cim. Suppl*. II, **8**, 532 (1958)

Weekes T.C. *Phys. Rep.* **160**, 1 (1998)

Weinberg S. *Phys. Rev. Lett.* **19**, 1264 (1967)

Wolfenstein L. *Phys. Rev.* **D17**, 2369 (1978)

Wu C.S. and I. Shaknov *Phys. Rev*. **77**, 136 (1950)

Yanagida T. *Physica Scripta* **T121**, 137 (2005)

Yao W.M. *et al. J. of Phys*. G **33**, 1 (2006), from Particle Data Group *Review of Particle Physics*

Yukawa H. *Proc. Math. Soc. Japan* **17**, 48 (1935)

Zatsepin G.T. and V.A. Kuzmin, *JETP* **4**, 53 (1966)

Zioutas K. *et al.* , *Phys. Rev. Lett*. **94** 121301 (2005)

Zweig R. *CERN Report* 8419/Th. 412 (1964)

Bibliography

Books on astrophysics and cosmology at a similar level

"*Cauldrons in the Cosmos*" C.E. Rolfs and W.S. Rodney (University of Chicago Press 1988)

"*The Big Bang*" J. Silk (W.H. Freeman and Co, New York 1989)

"*Cosmology and Particle Astrophysics*" L. Bergstrom and A. Goobar (Springer, 2nd edition Praxis Publishing, Chichester 2004)

"*Cosmology*" M. Rowan-Robinson (Clarendon Press, Oxford 1996)

"*The Dynamic Cosmos*" M.S. Madsen (Chapman and Hall, London 1995)

"*Particle Astrophysics*" H.V. Klapdor-Kleingrothaus and K. Zuber (IOP Publishing Ltd, Bristol 2000)

Books on astrophysics and cosmology at a more advanced level

"*The Early Universe*" E.W. Kolb and M.S. Turner (Addison-Wesley, New York 1990)

"*Principles of Physical Cosmology*" P.J.E. Peebles (Princeton University Press, Princeton, NJ 1993)

"*Introduction to Cosmology*" J.V. Narlikar (Cambridge University Press 2nd edition, Cambridge 1993)

"*Cosmological Physics*" J.A. Peacock (Cambridge University Press Cambridge 1999)

"*Introduction to Cosmology*" M. Roos (John Wiley and Sons, Chichester, 2nd edition 2004)

Books and articles on more specialist topics

"*General Relativity*" I.R. Kenyon (Oxford Science Publications, Oxford 1990)

"*Neutrino Astrophysics*" J. Bahcall (Cambridge University Press, Cambridge 1989)

"*The Physics of Stars*" A.C. Phillips (Manchester Physics Series, John Wiley and Sons, Chichester 1994)

"*Cosmic Rays and Particle Physics*" T.K. Gaisser (Cambridge University Press, Cambridge 1990)

"*Pulsar Astronomy*" A.G. Lyne and F. Graham-Smith (Cambridge Astrophysics Series, Cambridge University Press, Cambridge 1990)

"Very High Energy Gamma Ray Astronomy" R.A. Ong, *Phys. Rep.* **305**, 93–202 (1998)

"Very High Energy Gamma Ray Astronomy" T.C. Weekes, *Phys. Rep.* **160**, 1–121 (1988)

"Review of Gravitational Wave Detectors" F. Ricci and A. Brillet, *Ann. Rev. Nucl. Part. Sci.* **47**, 111 (1997)

"Nuclear Reactions in Stars" B.W. Filippone, *Ann. Rev. Nucl. Part. Sci.* **36**, 717 (1986)

"Neutrinos from Supernova Explosions" A. Burrows, *Ann. Rev. Nucl. Part. Sci.* **40**, 181 (1990)

"Search for Discrete Astrophysical Sources of Energetic Gamma Radiation" J.W. Cronin, K.G. Gibbs, and T.C. Weekes, *Ann.Rev. Nucl. Part. Sci.* **43**, 687 (1993)

"Gamma-Ray Bursts: Ligo/Virgo Sources of Gravitational Radiation" M. Van Putten, *Phys. Rep.* **345**, 1 (2001)

"*Cosmological Inflation and Large Scale Structure*" A.R. Liddle and D.H. Lyth (Cambridge University Press, Cambridge 2000)

"*High Energy Astrophysics*" M.S. Longair (Cambridge University Press, Cambridge, 2nd edition 1992)

"*Neutrino Physics*" K. Zuber (Institute of Physics Publishing, Bristol 2004)

Index

Abelian group, 74
Abell 2218 cluster, 161
absolute magnitude, 110
accelerating/decelerating universe, 126
accelerators, fixed target or colliding beam, 57
accretion disc, 256
acoustic peaks in microwave fluctuations, 216 *et seq*
action, and Hamilton's principle, 60
active galactic nuclei, see AGN
adiabatic fluctuations, 215
AGASA air shower experiment, 251
age of universe, 123
AGNs (active galactic nuclei), 255
AMANDA experiment, 265
Andromeda galaxy (M81), 108
angular momentum operators, 62
anisotropies in cosmic microwave radiation,
  primary and secondary anisotropies, 214 *et seq*
  experimental observations, 222
  angular spectrum, 216
  and temperature fluctuations, 213
anomalous magnetic moments, of electron and muon, 73
anthropic principle, 185
antineutrino,
  helicity, 67
antiparticles, 9
antiproton annihilation, 146
antishielding in QCD, 79
Arecibo radio telescope, 254
asymptotic freedom, 79
atmospheric neutrinos, 257
AUGER air shower experiment, 251
axions, 168
axion-like particles (ALPs), 169

B^0 meson decay and CP violation, 88
b quark (bottom quark), 4
Baksan experiment, 288
barn, millibarn, microbarn, etc, 22
baryogenesis, 146 *et seq*
baryons as quark combinations, 7
baryon-antibaryon asymmetry, 146 *et seq*
baryon-photon ratio, in Big Bang, 148
baryonic matter density, in universe, 121
baryonic dark matter, 166
BATSE experiment, 249
BEBC bubble chamber, 82
BeppoSAX satellite, 249
beta decay, and parity violation, 66
Bethe-Bloch energy loss formula, 240
Big Bang model, 112
Big Crunch, 116
binary pulsar, 267
binding energy, of nuclei, 274
Birkhoff theorem, 115
black body spectrum, 128
black holes, 294
  Hawking radiation from, 295
blazars, 257
Bohr magneton, 73
Boltzmann factor, 137
BOOMERANG experiment, 222
Bose-Einstein statistics, 9
Bose-Einstein distribution, 129
boson exchange, 11
boson propagator, 14
bottom (b) quark, 4
Breit-Wigner formula, 31
bremsstrahlung, 242
Brookhaven National Laboratory, 21

Cabibbo angle, 87
calorimeters, 28, 247
Carbon-12 resonance, 34
cascade showers, 243
Casimir effect, 182
CAST (CERN axion solar telescope), 169
CDF detector, 28
CDMS experiment, 176
centre-of-momentum frame (CMS), 23
Cepheid variables, 111
Cerenkov radiation, 241,
CfA catalogue, 109
Chandra X-ray satellite, 162
Chandrasekhar limit and mass, 284
chaotic inflation, 196
charge conjugation, and invariance, 69
charge conservation, 72
charm quark, 4
charmed meson, *see* J/ψ
clocks in gravitational fields, 42
CKM matrix, 87
cloud chamber, 11
closed universe, 116
closure parameters, of universe, 122
CMB, see cosmic microwave radiation,
CNGS experiment, 260
coasting universe, 181
COBE experiment, 127
cold dark matter, 207
collapsars, 249
colliding beam accelerators, 57
collision cross-section, definition , 22
  formula in 2-body → 2-body collision, 24
  spin factors in, 24
colour quantum number, 14
  in e^+e^-annihilation, 27
Coma galaxy cluster, 109
co-moving coordinates, in expanding universe, 112
Compton scattering, 247
confinement, of quarks, 7
conservation rules, and symmetries, 60 *et seq*
constituent quark masses, 7
continuity equation, 303
coordinate distance, 112
coordinate transformations, in relativity, 39
conversion length, 242
correlation function, in CMB anisotropy, 201
cosmic microwave radiation (CMB)
  anisotropy, dipole, 130
  anisotropies, multipole, 216–222
  discovery and spectrum, 127
  energy density of, 121,
  polarization, 223
  temperature fluctuations, 213
cosmic parameters, table of, 223
cosmic rays
  acceleration, 237
  abundance of elements in, 230
  east-west effect, 235
  geomagnetic effects on, 233
  hard and soft components, 239
  primary energy spectrum, 232
  showers, 243
  point sources, 253
cosmological constant, 118, 127
  and dark energy, 184
cosmological principle, 112

Coulomb barrier height, 274
Coulomb potential, 15
Coulomb scattering, 241
coupling constants, 8
covariant derivative, 71
CP symmetry, 83
CP eigenstates in neutral kaon system, 85
CP violation in neutral K and B meson decay, 84 *et seq*
CP violation and baryon asymmetry, 150
CP violation in Standard Model, 87
CPT theorem, 83, 151
Crab pulsar, 292
Crab gamma ray spectrum, 248
critical density of universe, 120
critical energy, for electrons, 242
cross-section, definition, 22
  examples of elementary cross-sections, 24
  and decay rates, 30
  for e^+e^- to anything, 27
current conservation, 71
curvature of space in general relativity, 20, 221
curvature parameter, 114
cycloid solution of closed universe, 117
Cygnus A radio galaxy, 256

dark energy
  density, 122
  evidence from high redshift supernovae, 176 *et seq*
  and the cosmological constant, 184
dark matter, 122
  in galaxies, 157
dark matter from gravitational lensing, 160
DAMA experiment, 176
DASI experiment, 222
decay rates and resonances, 32
$\Delta(1232)$ resonance, 33
  and GZK effect, 252
deceleration parameter, 126
decoupling of matter and radiation, 135
decoupling (or recombination) temperature, 137
degeneracy pressure, of electrons, 281
degrees of freedom in early universe, 133
density of final states, *see* phase space factor,
DESY laboratory, 17
deuterium production, in early universe, 144 *et seq*
diffraction scattering, 32
dipole anisotropy, of microwave radiation, 130
Dirac particles, 5
Dirac and Majorana mass matrices, 101
distances in cosmology
  true coordinate distance, 112, 126
  luminosity distance, 110
distance modulus, 110
distance – redshift relation, 108
divergences, in quantum field theory, 73
Doppler effect, transverse and longitudinal, 56
Doppler shift, 108
double beta decay, 92
  neutrinoless, 93
double pulsar, 268
down quark (d quark), 4

EDELWEISS experiment, 176
EGRET detector, 247
Einstein tensor and Einstein field equations, 48
Einstein ring, 160
Einstein star shift, 51
Einstein-de Sitter universe, 180
electric dipole moment, 91
electromagnetic coupling, 8
electromagnetic interactions, 14
electromagnetic showers, 243
electron,
  degeneracy pressure of, 281
  magnetic moment, 73
  radiation loss by, 242
electron neutrino, 4
electron-positron annihilation, 25 *et seq*
electron-positron colliders, 33, 58
electroweak interactions, 18
  and unification, 75
electroweak couplings, 75
emulsions, nuclear, 146
energy densities in universe, 119 *et seq*
energy-momentum tensor, 48
equation of state, 119
equivalence principle, 42 *et seq*
eta (η) meson, 91
Euler-Lagrange equation, 60
Euler's equation in fluid mechanics, 303
exchange interaction, 11
exclusion principle, 9
expansion parameter, of universe, 112
extensive air showers, 245

f-meson, 90
false vacuum, 76
Faraday rotation, 189
Fermi coupling constant, 17
Fermi's Second Golden Rule, 22
fermi unit (femtometer), 3
fermions,
  pair creation, 11
  parity of fermions and antifermions, 65,
  symmetry under interchange, 63
Fermi energy, 281
Fermi-Dirac statistics, 9
Fermi-Dirac distribution, 129, 131, 288
Fermilab, 28
Feynman diagrams, 12
fine structure constant, 14
fixed target experiments, 57
flat universe, 116
flatness problem, 190
flavour quantum number, 5
fluctuations
  in microwave radiation, 213 *et seq*
  adiabatic, isothermal or isocurvature, 208
  spectrum of, in CMB radiation, 207
  evolution during radiation era, 206
  growth during radiation-dominated era, 212
fluorescence, from atmospheric showers, 243
Fourier transform, 201
four-momentum transfer, 56
four vectors, 41
free-fall frame, as inertial frame, 46
free-fall time, 202
freeze-out condition, 135, 200
Friedmann equation, 114
Friedmann-Lemaitre-Robertson-Walker (FLRW) model, 54 , 114
fundamental interactions, table of, 8, 12
fusion probability and timescales, 281

g-factor, for electron and muon, 72
galaxies,
  spiral structure, 107
  galaxy clusters, 109
  irregular galaxies, 107
  magnetic fields in, 189
gamma rays
  absorption in matter, 242 *et seq*
  bursts, 249
  Compton scattering, 29
  conversion length, 242
  cosmic point sources, 247
gamma ray bursters, 293
Gamow energy, 275
Gargamelle neutrino experiment, 81
gauge bosons, 17
gauge invariance
  and gauge transformations, 69
  in electroweak theory, 74
gauge theories, vacuum structure in, 83
generalised coordinates, 61
general relativity, 47 *et seq*
GEO600 gravitational wave detector, 270
geodesics, 115
geomagnetic field, 233
GZK cut-off, 251
global phase transformations, 69
globular clusters, 278
gluons, 17
  colour charge, 15
  role in QCD, 81
g-factors of electron and muon, 73
grand unified theories (GUTs),
  baryogenesis in, 146
  couplings in, 99

proton decay in, 98
unified coupling, 97
supersymmetric version, 99
gravitation constant, 18
gravitational deflection of light, 51
gravitational interactions, 18
gravitational blue shift, 47
gravitational collapse, Jeans mass
and free-fall time, 202
gravitational lensing, 159
amplification by, 163
Einstein ring in, 160
microlensing, 163
gravitational pressure, 282
gravitational radiation, 264 *et seq*
and binary pulsar, 267
laboratory experiments, 268
quadrupole formula, 266
graviton, 8
gravitino, 154
GRO satellite, 247
GUTs, *see* grand unification
gyromagnetic ratio *see* g-factor
GZK cut-off, 251

hadrons, as quark combinations, 7
halos of galaxies, 157
Hamilton's principle of least action, 60
harmonic oscillator, 184
Hawking radiation, 295
heavy primary nuclei, in cosmic rays, 147
heavy element production, in stars, 280
helicity
definition and conservation, 67 *et seq*
helium-hydrogen ratio, in early universe, 143
helium burning, 279
heavy elements, stellar production, 280
HERA ep collider, 58
Herzsprung-Russell diagram, 278
hierarchy problem, 102
Higgs mechanism, 75
High z SN Search Team, 179
HiRes air shower experiment, 251
HOMESTAKE experiment, 261
horizon distance, optical, 126
horizon, acoustic, 216
horizon problem, 190
horizontal branch, 278
Hubble constant, 108
Hubble expansion, 107 *et seq*
Hubble plots, 111
at large redshifts, 178
for accelerating or decelerating universe, 180
Hubble space telescope, 111
Hubble time, 112
hydrostatic equilibrium, in stars, 273

ICECUBE neutrino experiment, 265
IMB experiment, 288
inelastic scattering in parton model, 79
inertial frame, 39
inflation, 192 *et seq*
chaotic inflation, 196
quantum fluctuations and inflation, 198
fluctuation spectrum, 200
slow roll approximation, 197
inflaton, 192
instantons, 153
interactions, table of, 8
interaction cross-section, definition and formula, 21 *et seq*
intermediate vector boson *see* W,Z bosons
intrinsic parity, 63
invariant intervals, in relativity, 41
inverse square law,
possible modifications, 55
ionization energy loss, 240
ionization potential, 241
IRAS galaxy survey, 207
isocurvature fluctuations, 208
isothermal fluctuations, 208
isospin, 8

J/ψ particle, 26
JADE detector, 17
Jeans length and Jeans mass, 202 *et seq*
jets in e^+e^- annihilation, 17

Kamiokande experiment, 288
K2K experiment, 260
K-mesons (charged kaons)
K^+ and K^- mesons,
K^0 mesons (neutral kaons)
CP violation in decays, 84
decay modes and lifetimes, 85
K_L,K_S and K_1,K_2 eigenstates, 85
Kepler's Law, 53
kinematic transformations, in relativity, 56
Klein-Gordon wave equation, 301

Λ, QCD scale parameter, 81
Λ, cosmological constant, 118
Lagrange equation, *see* Euler-Lagrange equation
Lagrangian energy density, 76
Large Hadron Collider (LHC), 21
Large Magellanic Cloud, 289
large-scale structure and Jeans mass, 202
last scattering surface, 222, 226
Lawrence Berkeley Laboratory, 146
Legendre polynomial expansion, 217
lensing, gravitational, 159
lensing optical depth, 165
LEP collider, 33
leptogenesis, in early universe, 152
leptoquark boson, 151
leptons, table and masses, 4, 7
lifetime
and widths of hadrons, 30
equality and CPT theorem, 84
light deflection by point mass, 51
light element abundances
and nucleosynthesis, 142
and spallation, in cosmic rays, 232
lithium production, early universe, 144
LIGO experiment, 270
lobes, of radio galaxies, 256
longitudinal polarization in beta decay, 66
loop diagrams, 72, 79
Lorentz factor in radio jets, 257, 272
Lorentz transformations, 39 *et seq*
LSP (lightest supersymmetric particle), 103
luminosity distance, 110
versus redshift, for different models, 179
Lyman alpha forest, 257

M31 galaxy, 108
MACHOs, 163
Magellanic Cloud, 111
Magic array, 247
magnetic fields
in intergalactic space, 189
in galaxies, 189
in neutron stars, 293
stellar, 190
magnetic moments, of electron and muon, 73
magnetic dipole moment, 91
magnetic monopoles, 194
magnitude (astronomical), 110
main sequence, in Herzsprung-Russell diagram, 278
Majorana neutrinos, 101
mass matrix, 101
masses, 153
MSTO (main sequence turn-off), 278
mass equality, for particle and antiparticle, 84
mass, gravitational and inertial, 45
mass generation, in electroweak theory, 75
matrix element, 22
matter-antimatter asymmetry, 146
matter-dominated universe,
growth of fluctuations in, 212
matter oscillations of neutrinos, 263
matter-radiation equality, 138
Mattig formula, 219
MAXIMA experiment, 222
Maxwell velocity distribution, 276
mean free path, 24
mesons, as quark-antiquark combinations, 7
metric tensor, 48
microlensing, and MACHOs, 163
microwave background, *see* cosmic microwave radiation
Milky Way, 108
Minkowski metric, 48

MINOS experiment, 260
Moliere unit, in scattering, 244
monopole problem, 194
MSTO (on main sequence) 278
MSW mechanism, 308
muon,
  decay rate, 30
  decay in atmosphere, 239
  flux in atmosphere, 104
muon-neutrino, 7

negative energy states, 10
negative pressure, 120
neutral weak currents, 17, 75
neutral K-meson, *see* K^0 mesons
neutral pion,
  and soft component of cosmic rays, 240
neutrinos
  as dark matter, 167
  atmospheric neutrinos, 96, 259
  cosmological mass limits, 210
  decoupling in early universe, 134
  densities in early universe, 135
  Dirac or Majorana, 6
  experiments at accelerators and reactors, 260
  flavour number, 6
  helicity, 67
  interaction cross-sections, 21
  mass differences, 96
  mass eigenstates, 95
  matter oscillations, 263, 308
  muon, electron and tauon flavours, 4
  point sources of high energy neutrinos, 263
  solar neutrinos, 96, 260
  supernova neutrinos, 288
  supernova mass limits, 290
neutrino interactions, examples, 29
neutrino flavour oscillations, 95
neutrino matter oscillations, 308
neutrinoless double beta decay, 93
neutron
  neutron-proton ratio and nucleosynthesis, 143
  electric dipole moment, 168
  lifetime, 143
neutronisation, 286
neutron stars and pulsars, 291
Newtonian coupling constant, 48
Newton's inverse square law, possible deviations, 55
Newton's law of spheres, 141
Noether's theorem, 70
nuclear binding energy, 274
nuclear cascade, 245
nuclear emulsion, 146
nucleosynthesis and light elements, 142

Olbers' paradox, 113
omega hyperon, 36
Ω- parameter, of universe, 121 *et seq*,
open universe, 116
orbital precession, 53

pair production, by gamma rays, 242
parallax, 110
parity
  conservation, 63
  intrinsic, 64
  of particle and antiparticle, 64
  of pion, 64
  operation of, 62
  violation in weak interactions, 65
partial width, 32
particles and antiparticles, 9
particle-antiparticle conjugation, 69
particle horizon, 191
particle and radiation densities in early universe, 131
parton model, in strong interactions, 79
Peccei-Quinn symmetry, and axion, 168
Pauli principle, 9
peculiar velocity, 113
perihelion advance, 53
perturbation theory of structure growth, 303
phase-space factor, 23
photino *see* supersymmetry
photons
  and gauge invariance, 69
  mass limits, 72
  polarizatioj in positronium decay, 65
  and neutrino densities in early universe, 134
  role in evolution of early universe, 207
photosphere, of Sun, 275
physical constants, Table of, 299
PHZ (Peebles-Harrison-Zeldovich) spectrum, 201
pions (*see also* neutral pions)
  as quark-antiquark combinations, 7
  pion and muon decay in atmosphere, 239
  parity, 64
  pion-proton resonance (Δ (1232)), 33
Planck mass and length, 19
point sources
  of high energy gamma rays, 247
  of neutrinos, 263
  of cosmic ray protons, 253
Poisson's equation, 49
polarization
  of leptons in beta decay, 66
  of photons in positronium decay, 65
  of cosmic microwave radiation, 223
positron, discovery, 11
positronium
  decay modes and lifetimes, 64
power, in fluctuation spectrum, 202
pp cycle in Sun, 276
precession of orbits in general relativity, 53
primary anisotropies, in microwave background, 214
primordial abundances of elements, 144
primordial nucleosynthesis, 142
propagator term, 14
proton decay experiments, 100
proton decay, and radiation dosage,104
proton-antiproton annihilation, 146
  CP eigenstates in, 86
pseudoscalar mesons, 91
protostars, 273
pulsars, 291

quadrupole moment, 266
quantum chromodynamics, 78
quantum electrodynamics, 73
quantum exchange, 11
quantum fluctuations (in inflation), 198
quantum gravity, 73
quark confinement, 15
quark flavours, 6
quark masses, 7
quark model, 34
quark-gluon plasma, 21
quark flavour mixing in weak interactions, 87
quark-parton model, of deep inelastic scattering, 79
quarks, table of 4, 7
quark interactions, 15
quasars, 253
  and Lyman alpha forest, 257
  and lensing double image, 161
quintessence, 182

radiation and matter eras, 135
radiation-dominated universe, 135
radiation dosage, 104
radiation length, 242
radiation loss of electrons, 242
radiative corrections,
  in quantum electrodynamics, 72
  in quantum chromodynamics, 79
radiative transfer in stars, 275
radio galaxies, 253
radio lobes, 254
radio telescopes, 254
radioactive decay, 30
radius of universe, 125
reactor neutrino experiments, 97
recombination era, 135
red giant branch, 278
redshift
  cosmological, 108 *et seq*
reheating, after inflation, 198
relativistic transformations, 39 *et seq*
renormalization, in field theory, 72
resonances, and Breit-Wigner formula, 30 *et seq*
resonances in astrophysics, 34
RHIC, 21

Ricci tensor and scalar, 48
Riemann curvature tensor, 48
rho (ρ) meson, 90, 91
Robertson-Walker line element, 54
rotation curves of galaxies, 157 *et seq*
rotation operator, 62
running couplings, 81 *et seq*
Rutherford scattering, 25

Sachs-Wolfe effect, 214
Sagiittarius A*, 298
Saha equation, 137
Sakharov criteria, 150
scale parameter, 116
scale-invariant fluctuation spectrum, 201
scattering amplitude, 31
Schwarzschild line element, 49
Schwarzschild radius, 49
  and black holes, 50, 294
secondary anisotropies, in microwave background, 215
see-saw mechanism, 101
Shapiro time delay, 52
shielding effects in field theory, 78
shock front in cosmic ray acceleration, 237
shower development, 243
sigma hyperons, 37
Silk damping, 208
SLAC laboratory, 81
Sloane Digital Sky Survey, 211
SN 1987A, 288
  light curve, 291
SNO experiment, 262
solar neutrinos, 260
spark chamber, 5, 247
special relativity, 39 *et seq*
sphalerons, 153
spherical harmonics, 63
spin functions, for two fermions, 63
spin-statistics relation, 9
spiral galaxies, 108
spontaneous symmetry breaking, 76
Standard Model, of particle physics, 60 *et seq*
steady state theory, 112
stellar evolution and equilibrium, 273 *et seq*
stellar collapse, 281
strange particles, strangeness, 8
string model, of gluons, 15
strong CP problem, 168
strong interactions, and colour, 15
structure functions, in inelastic scattering, 80
structure growth, in expanding universe, 205
Sunyaev-Zeldovich effect, 216
Superkamiokande experiment, 100, 259, 261
  detector, 100
superluminal expansion, 193
supernovae
  1987A, 111, 288
  in Hubble plots, 174 *et seq*
  light curve, 291
  neutrinos from, 289
  high redshift type Ia, 177
  thermal phase, 287
  type II, 285
Supernova Cosmology Project, 179
superstring theory, 73
supersymmetry (SUSY), 9 , 103
  and grand unification, 99
  and hierarchy problem, 102
  table of SUSY particles, 103
  lightest SUSY particle, and dark matter, 103
symmetry groups in particle physics,
  U(1), 74
  SU(2), 74
  SU(3), 97
  SU(5), 97, 151
  SO(10), 152
synchrotron radiation, 256

2 degree Field Galaxy Redshift Survey, 211
TAMA gravity wave detector, 270
tau lepton and tau neutrino, 4
temperature fluctuations, in CMB, 213,
tensor field, in gravitation, 47
Tevatron, Fermilab, 28
tidal forces, 43
tilt, 202
time reversal(T) invariance,
thermonuclear reactions, in stars, 274 *et seq*
Thomson cross-section, 29
Thomson scattering, 213 - 216
  and polarisation of CMB, 223
Thomson drag, 208
three-jet events, 17
tidal forces, 43
  lunar and solar tides, 44
time reversal (T) symmetry, 83, 91
top quark,
  discovery, 28
  mass from Standard Model fits, 82
  production rate and properties, 29
track length integral, 244
transition matrix, 25
transverse momentum, 245
triangle anomaly, 16
triple alpha process, 279
true vacuum., 76
turn-off point on main sequence (MSTO),

uncertainty principle, 13
unification of electromagnetic and weak interactions, 18
unification energy, in GUTs, 99
unitarity condition, in CKM matrix, 88
units in particle physics, 4
unitarity triangle, in B-meson decay, 89
universe, approximate masses in, 110
  age and size, 123
up quark (u quark), 4
upsilon meson, 91

vacuum energy, 119, 122
  and Casimir effect, 182
  effect on deceleration parameter, 126
  and quintessence, 182
vacuum expectation value, 76
vacuum polarization, 78, 82
vacuum structure, in gauge theories, 83
vector interactions, and helicity conservation,
VERITAS mirror array, 246
VIRGO gravity wave detector, 270
virial theorem, 157
virtual particle exchange, 13
Very Large Telescope (ESO), 162
VLA radio telescope, 255

w-parameter, for dark energy, 122
W boson, discovery, 18
W and Z boson mediators, 17
  masses in electroweak theory, 75
  width of Z boson, 33
weak interactions and weak coupling, 17
weak mixing angle, 75
weak neutral currents, *see* neutral currents,
weakly interacting massive particles, *see* WIMPs
WHIPPLE observatory, 246
white dwarfs
  and degeneracy pressure, 284
  cooling rates, 285
width of resonance, 30
WIMPs,
  Interaction and annihilation cross-sections, 172
  experimental searches, 174
Wilkinson Microwave Anisotropy Probe (WMAP), 108, 216
Wolf-Rayet stars, 249

X,Y leptoquark bosons, 98
X-ray pulsars, X-ray bursters, 293
X-rays from galaxy clusters, 158

Yukawa coupling and potential, 13, 301
  propagator term, 14

Z^0 boson
  mass and width, 33
  resonance in e^+e^- annihilation, 33
ZEPLIN experiment, 176
zero point energy and vacuum energy density, 184