إلى قارئ هذا الكتاب ، تحية طيبة وبعد ...

لقد أصبحنا نعيش في عالم يعج بالأبحاث والكتب والمعلومات، وأصبح العلم معياراً حقيقياً لتفاضل الأمم والدول والمؤسسات والأشخاص على حدٍّ سواء، وقد أمسى بدوره حلاً شبه وحيدٍ لأكثر مشاكل العالم حدة وخطورة، فالبيئة تبحث عن حلول، وصحة الإنسان تبحث عن حلول، والموارد التي تشكل حاجة أساسية للإنسان تبحث عن حلول كذلك، والطاقة والغذاء والماء جميعها تحديات يقف العلم في وجهها الآن ويحاول أن يجد الحلول لها. فأين نحن من هذا العلم ؟ وأين هو منا؟

نسعى في موقع عالم الإلكترون www.4electron.com لأن نوفر بين أيدي كل من حمل على عاتقه مسيرة درب تملؤه التحديات ما نستطيع من أدوات تساعده في هذا الدرب، من مواضيع علمية، ومراجع أجنبية بأحدث إصداراتها، وساحات لتبادل الآراء والأفكار العلمية والمرتبطة بحياتنا الهندسية، وشروحٍ لأهم برمجيات الحاسب التي تتداخل مع تطبيقات الحياة الأكاديمية والعملية، ولكننا نتوقع في نفس الوقت أن نجد بين الطلاب والمهندسين والباحثين من يسعى مثلنا لتحقيق النفع والفائدة للجميع، ويحلم أن يكون عضواً في مجتمعٍ يساهم بتحقيق بيئة خصبة للمواهب والإبداعات والتألق، فهل تحلم بذلك ؟

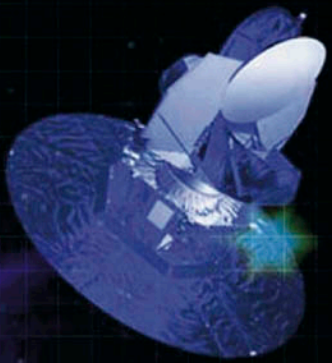
حاول أن تساهم بفكرة، بومضة من خواطر تفكيرك العلمي، بفائدة رأيتها في إحدى المواضيع العلمية، بجانب مضيء لمحته خلف ثنايا مفهوم هندسي ما. تأكد بأنك ستلتمس الفائدة في كل خطوة تخطوها، وترى غيرك يخطوها معك ...

أخي القارئ، نرجو أن يكون هذا الكتاب مقدمة لمشاركتك في عالمنا العلمي التعاوني، وسيكون موقعكم عالم الإلكترون ww.4electron.com بكل الإمكانيات المتوفرة لديه جاهزاً على الدوام لأن يحقق البيئة والواقع الذي يبحث عنه كل باحث أو طالب في علوم الهندسة، ويسعى فيه للإفادة كل ساعٍ ، فأهلاً وسهلاً بكم .

مع تحيات إدارة الموقع وفريق عمله

www.4electron.com

Matts Roos



Introduction to Cosmology

Third Edition

WILEY

Introduction to Cosmology

Third Edition

Matts Roos

John Wiley & Sons, Ltd

Introduction to Cosmology

Third Edition

Introduction to Cosmology

Third Edition

Matts Roos

John Wiley & Sons, Ltd

Copyright © 2003 John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk

Visit our Home Page on www.wileyeurope.com or www.wiley.com

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, electronic, mechanical, photocopying, recording, scanning or otherwise, except under the terms of the Copyright, Designs and Patents Act 1988 or under the terms of a licence issued by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London W1T 4LP, UK, without the permission in writing of the Publisher. Requests to the Publisher should be addressed to the Permissions Department, John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, or emailed to permreq@wiley.co.uk, or faxed to (+44) 1243 770571.

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the Publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 33 Park Road, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Library of Congress Cataloging-in-Publication Data

Roos, Matts.
Introduction to cosmology / Matt Roos. – 3rd ed.
p. cm.
Includes bibliographical references and index.
ISBN 0-470-84909-6 (acid-free paper) – ISBN 0-470-84910-X (pbk. : acid-free paper)
1. Cosmology. I. Title.

QB981.R653 2003
523.1 — dc22

2003020688

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0 470 84909 6 (hardback)
0 470 84910 X (paperback)

Typeset in 9.5/12.5pt Lucida Bright by T&T Productions Ltd, London.
Printed and bound in Great Britain by Antony Rowe Ltd., Chippenham, Wilts.
This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

To my dear grandchildren

Francis Alexandre Wei Ming (1986)
Christian Philippe Wei Sing (1990)

Cornelia (1989)
Erik (1991)
Adrian (1994)

Emile Johannes (2000)
Alaia Ingrid Markuntytär (2002)

Contents

Introduction to Cosmology Third Edition by Matts Roos
© 2003 John Wiley & Sons, Ltd ISBN 0 470 84909 6 (cased) ISBN 0 470 84910 X (pbk)

Preface to First Edition

A few decades ago, astronomy and particle physics started to merge in the common field of cosmology. The general public had always been more interested in the visible objects of astronomy than in invisible atoms, and probably met cosmology first in Steven Weinberg's famous book *The First Three Minutes*. More recently Stephen Hawking's *A Brief History of Time* has caused an avalanche of interest in this subject.

Although there are now many popular monographs on cosmology, there are so far no introductory textbooks at university undergraduate level. Chapters on cosmology can be found in introductory books on relativity or astronomy, but they cover only part of the subject. One reason may be that cosmology is explicitly cross-disciplinary, and therefore it does not occupy a prominent position in either physics or astronomy curricula.

At the University of Helsinki I decided to try to take advantage of the great interest in cosmology among the younger students, offering them a one-semester course about one year before their specialization started. Hence I could not count on much familiarity with quantum mechanics, general relativity, particle physics, astrophysics or statistical mechanics. At this level, there are courses with the generic name of Structure of Matter dealing with Lorentz transformations and the basic concepts of quantum mechanics. My course aimed at the same level. Its main constraint was that it had to be taught as a one-semester course, so that it would be accepted in physics and astronomy curricula. The present book is based on that course, given three times to physics and astronomy students in Helsinki.

Of course there already exist good books on cosmology. The reader will in fact find many references to such books, which have been an invaluable source of information to me. The problem is only that they address a postgraduate audience that intends to specialize in cosmology research. My readers will have to turn to these books later when they have mastered all the professional skills of physics and mathematics.

In this book I am not attempting to teach basic physics to astronomers. They will need much more. I am trying to teach just enough physics to be able to explain the main ideas in cosmology without too much hand-waving. I have tried to avoid the other extreme, practised by some of my particle physics colleagues, of writing books on cosmology with the obvious intent of making particle physicists out of every theoretical astronomer.

Introduction to Cosmology Third Edition by Matts Roos
© 2003 John Wiley & Sons, Ltd ISBN 0 470 84909 6 (cased) ISBN 0 470 84910 X (pbk)

I also do not attempt to teach basic astronomy to physicists. In contrast to astronomy scholars, I think the main ideas in cosmology do not require very detailed knowledge of astrophysics or observational techniques. Whole books have been written on distance measurements and the value of the Hubble parameter, which still remains imprecise to a factor of two. Physicists only need to know that quantities entering formulae are measurable—albeit incorporating factors h to some power—so that the laws can be discussed meaningfully. At undergraduate level, it is not even usual to give the errors on measured values.

In most chapters there are subjects demanding such a mastery of theoretical physics or astrophysics that the explanations have to be qualitative and the derivations meagre, for instance in general relativity, spontaneous symmetry breaking, inflation and galaxy formation. This is unavoidable because it just reflects the level of undergraduates. My intention is to go just a few steps further in these matters than do the popular monographs.

I am indebted in particular to two colleagues and friends who offered constructive criticism and made useful suggestions. The particle physicist Professor Kari Enqvist of NORDITA, Copenhagen, my former student, has gone to the trouble of reading the whole manuscript. The space astronomer Professor Stuart Bowyer of the University of California, Berkeley, has passed several early mornings of jet lag in Lapland going through the astronomy-related sections. Anyway, he could not go out skiing then because it was either a snow storm or -30 °C! Finally, the publisher provided me with a very knowledgeable and thorough referee, an astrophysicist no doubt, whose criticism of the chapter on galaxy formation was very valuable to me. For all remaining mistakes I take full responsibility. They may well have been introduced by me afterwards.

Thanks are also due to friends among the local experts: particle physicist Professor Masud Chaichian and astronomer Professor Kalevi Mattila have helped me with details and have answered my questions on several occasions. I am also indebted to several people who helped me to assemble the pictorial material: Drs Subir Sarkar in Oxford, Rocky Kolb in the Fermilab, Carlos Frenk in Durham, Werner Kienzle at CERN and members of the COBE team.

Finally, I must thank my wife Jacqueline for putting up with almost two years of near absence and full absent-mindedness while writing this book.

Matts Roos

Preface to Second Edition

In the three years since the first edition of this book was finalized, the field of cosmology has seen many important developments, mainly due to new observations with superior instruments such as the Hubble Space Telescope and the ground-based Keck telescope and many others. Thus a second edition has become necessary in order to provide students and other readers with a useful and up-to-date textbook and reference book.

At the same time I could balance the presentation with material which was not adequately covered before—there I am in debt to many readers. Also, the inevitable number of misprints, errors and unclear formulations, typical of a first edition, could be corrected. I am especially indebted to Kimmo Kainulainen who served as my course assistant one semester, and who worked through the book and the problems thoroughly, resulting in a very long list of corrigenda. A similar shorter list was also dressed by George Smoot and a student of his. It still worries me that the errors found by George had been found neither by Kimmo nor by myself, thus statistics tells me that some errors still will remain undetected.

For new pictorial material I am indebted to Wes Colley at Princeton, Carlos Frenk in Durham, Charles Lineweaver in Strasbourg, Jukka Nevalainen in Helsinki, Subir Sarkar in Oxford, and George Smoot in Berkeley. I am thankful to the Academie des Sciences for an invitation to Paris where I could visit the Observatory of Paris-Meudon and profit from discussions with S. Bonazzola and Brandon Carter.

Several of my students have contributed in various ways: by misunderstandings, indicating the need for better explanations, by their enthusiasm for the subject, and by technical help, in particular S. M. Harun-or-Rashid. My youngest grandchild Adrian (not yet 3) has showed a vivid interest for supernova bangs, as demonstrated by an X-ray image of the Cassiopeia A remnant. Thus the future of the subject is bright.

Matts Roos

Introduction to Cosmology Third Edition by Matts Roos
© 2003 John Wiley & Sons, Ltd ISBN 0 470 84909 6 (cased) ISBN 0 470 84910 X (pbk)

Preface to Third Edition

This preface can start just like the previous one: in the seven years since the second edition was finalized, the field of cosmology has seen many important developments, mainly due to new observations with superior instruments. In the past, cosmology often relied on philosophical or aesthetic arguments; now it is maturing to become an exact science. For example, the Einstein–de Sitter universe, which has zero cosmological constant ($\Omega_\lambda = 0$), used to be favoured for esthetical reasons, but today it is known to be very different from zero ($\Omega_\lambda = 0.73 \pm 0.04$).

In the first edition I quoted $\Omega_0 = 0.8 \pm 0.3$ (daring to believe in errors that many others did not), which gave room for all possible spatial geometries: spherical, flat and hyperbolic. Since then the value has converged to $\Omega_0 = 1.02 \pm 0.02$, and everybody is now willing to concede that the geometry of the Universe is flat, $\Omega_0 = 1$. This result is one of the cornerstones of what we now can call the 'Standard Model of Cosmology'. Still, deep problems remain, so deep that even Einstein's general relativity is occasionally put in doubt.

A consequence of the successful march towards a 'standard model' is that many alternative models can be discarded. An introductory text of limited length like the current one cannot be a historical record of failed models. Thus I no longer discuss, or discuss only briefly, $k \neq 0$ geometries, the Einstein–de Sitter universe, hot and warm dark matter, cold dark matter models with $\Lambda = 0$, isocurvature fluctuations, topological defects (except monopoles), Bianchi universes, and formulae which only work in discarded or idealized models, like Mattig's relation and the Saha equation.

Instead, this edition contains many new or considerably expanded subjects: Section 2.3 on Relativistic Distance Measures, Section 3.3 on Gravitational Lensing, Section 3.5 on Gravitational Waves, Section 4.3 on Dark Energy and Quintessence, Section 5.1 on Photon Polarization, Section 7.4 on The Inflaton as Quintessence, Section 7.5 on Cyclic Models, Section 8.3 on CMB Polarization Anisotropies, Section 8.4 on model testing and parameter estimation using mainly the first-year CMB results of the Wilkinson Microwave Anisotropy Probe, and Section 9.5 on large-scale structure results from the 2 degree Field (2dF) Galaxy Redshift Survey. The synopsis in this edition is also different and hopefully more logical, much has been entirely rewritten, and all parameter values have been updated.

I have not wanted to go into pure astrophysics, but the line between cosmology and cosmologically important astrophysics is not easy to draw. Supernova explosion mechanisms and black holes are included as in the earlier editions, but not

Introduction to Cosmology Third Edition by Matts Roos
© 2003 John Wiley & Sons, Ltd ISBN 0 470 84909 6 (cased) ISBN 0 470 84910 X (pbk)

for instance active galactic nuclei (AGNs) or jets or ultra-high-energy cosmic rays. Observational techniques are mentioned only briefly—they are beyond the scope of this book.

There are many new figures for which I am in debt to colleagues and friends, all acknowledged in the figure legends. I have profited from discussions with Professor Carlos Frenk at the University of Durham and Professor Kari Enqvist at the University of Helsinki. I am also indebted to Professor Juhani Keinonen at the University of Helsinki for having generously provided me with working space and access to all the facilities at the Department of Physical Sciences, despite the fact that I am retired.

Many critics, referees and other readers have made useful comments that I have tried to take into account. One careful reader, Urbana Lopes França Jr, sent me a long list of misprints and errors. A critic of the second edition stated that the errors in the first edition had been corrected, but that new errors had emerged in the new text. This will unfortunately always be true in any comparison of edition $n + 1$ with edition n. In an attempt to make continuous corrections I have assigned a web site for a list of errors and misprints. The address is

http://www.physics.helsinki.fi/~fl_cosmo/

My most valuable collaborator has been Thomas S. Coleman, a nonphysicist who contacted me after having spotted some errors in the second edition, and who proposed some improvements in case I were writing a third edition. This came at the appropriate time and led to a collaboration in which Thomas S. Coleman read the whole manuscript, corrected misprints, improved my English, checked my calculations, designed new figures and proposed clarifications where he found the text difficult.

My wife Jacqueline has many interesting subjects of conversation at the breakfast table. Regretfully, her breakfast companion is absent-minded, thinking only of cosmology. I thank her heartily for her kind patience, promising improvement.

Matts Roos
Helsinki, March 2003

1

From Newton to Hubble

The history of ideas on the structure and origin of the Universe shows that humankind has always put itself at the centre of creation. As astronomical evidence has accumulated, these anthropocentric convictions have had to be abandoned one by one. From the natural idea that the solid Earth is at rest and the celestial objects all rotate around us, we have come to understand that we inhabit an average-sized planet orbiting an average-sized sun, that the Solar System is in the periphery of a rotating galaxy of average size, flying at hundreds of kilometres per second towards an unknown goal in an immense Universe, containing billions of similar galaxies.

Cosmology aims to explain the origin and evolution of the entire contents of the Universe, the underlying physical processes, and thereby to obtain a deeper understanding of the laws of physics assumed to hold throughout the Universe. Unfortunately, we have only one universe to study, the one we live in, and we cannot make experiments with it, only observations. This puts serious limits on what we can learn about the origin. If there are other universes we will never know.

Although the history of cosmology is long and fascinating, we shall not trace it in detail, nor any further back than Newton, accounting (in Section 1.1) only for those ideas which have fertilized modern cosmology directly, or which happened to be right although they failed to earn timely recognition. In the early days of cosmology, when little was known about the Universe, the field was really just a branch of philosophy.

Having a rigid Earth to stand on is a very valuable asset. How can we describe motion except in relation to a fixed point? Important understanding has come from the study of inertial systems, in uniform motion with respect to one another. From the work of Einstein on inertial systems, the theory of special relativity

Introduction to Cosmology Third Edition by Matts Roos
© 2003 John Wiley & Sons, Ltd ISBN 0 470 84909 6 (cased) ISBN 0 470 84910 X (pbk)

was born. In Section 1.2 we discuss inertial frames, and see how expansion and contraction are natural consequences of the homogeneity and isotropy of the Universe.

A classic problem is why the night sky is dark and not blazing like the disc of the Sun, as simple theory in the past would have it. In Section 1.3 we shall discuss this so-called Olbers' paradox, and the modern understanding of it.

The beginning of modern cosmology may be fixed at the publication in 1929 of Hubble's law, which was based on observations of the redshift of spectral lines from remote galaxies. This was subsequently interpreted as evidence for the expansion of the Universe, thus ruling out a static Universe and thereby setting the primary requirement on theory. This will be explained in Section 1.4. In Section 1.5 we turn to determinations of cosmic timescales and the implications of Hubble's law for our knowledge of the age of the Universe.

In Section 1.6 we describe Newton's theory of gravitation, which is the earliest explanation of a gravitational force. We shall 'modernize' it by introducing Hubble's law into it. In fact, we shall see that this leads to a cosmology which already contains many features of current Big Bang cosmologies.

1.1 Historical Cosmology

At the time of *Isaac Newton* (1642–1727) the heliocentric Universe of *Nicolaus Copernicus* (1473–1543), *Galileo Galilei* (1564–1642) and *Johannes Kepler* (1571–1630) had been accepted, because no sensible description of the motion of the planets could be found if the Earth was at rest at the centre of the Solar System. Humankind was thus dethroned to live on an average-sized planet orbiting around an average-sized sun.

The stars were understood to be suns like ours with fixed positions in a static Universe. The Milky Way had been resolved into an accumulation of faint stars with the telescope of Galileo. The *anthropocentric view* still persisted, however, in locating the Solar System at the centre of the Universe.

Newton's Cosmology. The first theory of gravitation appeared when Newton published his *Philosophiae Naturalis Principia Mathematica* in 1687. With this theory he could explain the empirical laws of Kepler: that the planets moved in elliptical orbits with the Sun at one of the focal points. An early success of this theory came when *Edmund Halley* (1656–1742) successfully predicted that the comet sighted in 1456, 1531, 1607 and 1682 would return in 1758. Actually, the first observation confirming the heliocentric theory came in 1727 when *James Bradley* (1693–1762) discovered the aberration of starlight, and explained it as due to the changes in the velocity of the Earth in its annual orbit. In our time, Newton's theory of gravitation still suffices to describe most of planetary and satellite mechanics, and it constitutes the nonrelativistic limit of Einstein's relativistic theory of gravitation.

Newton considered the stars to be suns evenly distributed throughout infinite space in spite of the obvious concentration of stars in the Milky Way. A distribution is called *homogeneous* if it is uniformly distributed, and it is called *isotropic* if it has the same properties in all spatial directions. Thus in a homogeneous and isotropic space the distribution of matter would look the same to observers located anywhere—no point would be preferential. Each local region of an isotropic universe contains information which remains true also on a global scale. Clearly, matter introduces lumpiness which grossly violates homogeneity on the scale of stars, but on some larger scale isotropy and homogeneity may still be a good approximation. Going one step further, one may postulate what is called the *cosmological principle*, or sometimes the *Copernican principle*.

> *The Universe is homogeneous and isotropic in three-dimensional space, has always been so, and will always remain so.*

It has always been debated whether this principle is true, and on what scale. On the galactic scale visible matter is lumpy, and on larger scales galaxies form gravitationally bound clusters and narrow strings separated by voids. But galaxies also appear to form loose groups of three to five or more galaxies. Several surveys have now reached agreement that the distribution of these galaxy groups appears to be homogeneous and isotropic within a sphere of 170 Mpc radius [1]. This is an order of magnitude larger than the supercluster to which our Galaxy and our local galaxy group belong, and which is centred in the constellation of Virgo.

Based on his theory of gravitation, Newton formulated a cosmology in 1691. Since all massive bodies attract each other, a finite system of stars distributed over a finite region of space should collapse under their mutual attraction. But this was not observed, in fact the stars were known to have had fixed positions since antiquity, and Newton sought a reason for this stability. He concluded, erroneously, that the self-gravitation within a finite system of stars would be compensated for by the attraction of a sufficient number of stars outside the system, distributed evenly throughout infinite space. However, the total number of stars could not be infinite because then their attraction would also be infinite, making the static Universe unstable. It was understood only much later that the addition of external layers of stars would have no influence on the dynamics of the interior. The right conclusion is that the Universe cannot be static, an idea which would have been too revolutionary at the time.

Newton's contemporary and competitor *Gottfried Wilhelm von Leibnitz* (1646–1716) also regarded the Universe to be spanned by an abstract infinite space, but in contrast to Newton he maintained that the stars must be infinite in number and distributed all over space, otherwise the Universe would be bounded and have a centre, contrary to contemporary philosophy. Finiteness was considered equivalent to boundedness, and infinity to unboundedness.

Rotating Galaxies. The first description of the Milky Way as a rotating galaxy can be traced to *Thomas Wright* (1711–1786), who wrote *An Original Theory or New Hypothesis of the Universe* in 1750, suggesting that the stars are

> *all moving the same way and not much deviating from the same plane, as the planets in their heliocentric motion do round the solar body.*

Wright's galactic picture had a direct impact on *Immanuel Kant* (1724–1804). In 1755 Kant went a step further, suggesting that the diffuse nebulae which Galileo had already observed could be distant galaxies rather than nearby clouds of incandescent gas. This implied that the Universe could be homogeneous on the scale of galactic distances in support of the cosmological principle.

Kant also pondered over the reason for transversal velocities such as the movement of the Moon. If the Milky Way was the outcome of a gaseous nebula contracting under Newton's law of gravitation, why was all movement not directed towards a common centre? Perhaps there also existed repulsive forces of gravitation which would scatter bodies onto trajectories other than radial ones, and perhaps such forces at large distances would compensate for the infinite attraction of an infinite number of stars? Note that the idea of a contracting gaseous nebula constituted the first example of a nonstatic system of stars, but at galactic scale with the Universe still static.

Kant thought that he had settled the argument between Newton and Leibnitz about the finiteness or infiniteness of the system of stars. He claimed that either type of system embedded in an infinite space could not be stable and homogeneous, and thus the question of infinity was irrelevant. Similar thoughts can be traced to the scholar *Yang Shen* in China at about the same time, then unknown to Western civilization [2].

The infinity argument was, however, not properly understood until *Bernhard Riemann* (1826–1866) pointed out that the world could be *finite* yet *unbounded*, provided the geometry of the space had a positive curvature, however small. On the basis of Riemann's geometry, *Albert Einstein* (1879–1955) subsequently established the connection between the geometry of space and the distribution of matter.

Kant's repulsive force would have produced trajectories in random directions, but all the planets and satellites in the Solar System exhibit transversal motion in one and the same direction. This was noticed by *Pierre Simon de Laplace* (1749–1827), who refuted Kant's hypothesis by a simple probabilistic argument in 1825: the observed movements were just too improbable if they were due to random scattering by a repulsive force. Laplace also showed that the large transversal velocities and their direction had their origin in the rotation of the primordial gaseous nebula and the law of conservation of angular momentum. Thus no repulsive force is needed to explain the transversal motion of the planets and their moons, no nebula could contract to a point, and the Moon would not be expected to fall down upon us.

This leads to the question of the origin of time: what was the first cause of the rotation of the nebula and when did it all start? This is the question modern cosmology attempts to answer by tracing the evolution of the Universe backwards in time and by reintroducing the idea of a repulsive force in the form of a cosmological constant needed for other purposes.

Black Holes. The implications of Newton's gravity were quite well understood by *John Michell* (1724–1793), who pointed out in 1783 that a sufficiently massive and compact star would have such a strong gravitational field that nothing could escape from its surface. Combining the corpuscular theory of light with Newton's theory, he found that a star with the solar density and escape velocity c would have a radius of $486R_\odot$ and a mass of 120 million solar masses. This was the first mention of a type of star much later to be called a *black hole* (to be discussed in Section 3.4). In 1796 Laplace independently presented the same idea.

Galactic and Extragalactic Astronomy. Newton should also be credited with the invention of the reflecting telescope—he even built one—but the first one of importance was built one century later by *William Herschel* (1738–1822). With this instrument, observational astronomy took a big leap forward: Herschel and his son John could map the nearby stars well enough in 1785 to conclude correctly that the Milky Way was a disc-shaped star system. They also concluded erroneously that the Solar System was at its centre, but many more observations were needed before it was corrected. Herschel made many important discoveries, among them the planet Uranus, and some 700 binary stars whose movements confirmed the validity of Newton's theory of gravitation outside the Solar System. He also observed some 250 diffuse nebulae, which he first believed were distant galaxies, but which he and many other astronomers later considered to be nearby incandescent gaseous clouds belonging to our Galaxy. The main problem was then to explain why they avoided the directions of the galactic disc, since they were evenly distributed in all other directions.

The view of Kant that the nebulae were distant galaxies was also defended by *Johann Heinrich Lambert* (1728–1777). He came to the conclusion that the Solar System along, with the other stars in our Galaxy, orbited around the galactic centre, thus departing from the heliocentric view. The correct reason for the absence of nebulae in the galactic plane was only given by *Richard Anthony Proctor* (1837–1888), who proposed the presence of interstellar dust. The arguments for or against the interpretation of nebulae as distant galaxies nevertheless raged throughout the 19th century because it was not understood how stars in galaxies more luminous than the whole galaxy could exist—these were observations of supernovae. Only in 1925 did *Edwin P. Hubble* (1889–1953) resolve the conflict indisputably by discovering Cepheids and ordinary stars in nebulae, and by determining the distance to several galaxies, among them the celebrated M31 galaxy in the *Andromeda*. Although this distance was off by a factor of two, the conclusion was qualitatively correct.

In spite of the work of Kant and Lambert, the heliocentric picture of the Galaxy—or almost heliocentric since the Sun was located quite close to Herschel's galactic centre—remained long into our century. A decisive change came with the observations in 1915–1919 by *Harlow Shapley* (1895–1972) of the distribution of *globular clusters* hosting 10^5–10^7 stars. He found that perpendicular to the galactic plane they were uniformly distributed, but along the plane these clusters had a distribution which peaked in the direction of the Sagittarius. This defined the centre

of the Galaxy to be quite far from the Solar System: we are at a distance of about two-thirds of the galactic radius. Thus the anthropocentric world picture received its second blow—and not the last one—if we count Copernicus's heliocentric picture as the first one. Note that Shapley still believed our Galaxy to be at the centre of the astronomical Universe.

The End of Newtonian Cosmology. In 1883 *Ernst Mach* (1838–1916) published a historical and critical analysis of mechanics in which he rejected Newton's concept of an absolute space, precisely because it was unobservable. Mach demanded that the laws of physics should be based only on concepts which could be related to observations. Since motion still had to be referred to some frame at rest, he proposed replacing absolute space by an idealized rigid frame of fixed stars. Thus 'uniform motion' was to be understood as motion relative to the whole Universe. Although Mach clearly realized that all motion is relative, it was left to Einstein to take the full step of studying the laws of physics as seen by observers in inertial frames in relative motion with respect to each other.

Einstein published his General Theory of Relativity in 1917, but the only solution he found to the highly nonlinear differential equations was that of a static Universe. This was not so unsatisfactory though, because the then known Universe comprised only the stars in our Galaxy, which indeed was seen as static, and some nebulae of ill-known distance and controversial nature. Einstein firmly believed in a static Universe until he met Hubble in 1929 and was overwhelmed by the evidence for what was to be called Hubble's law.

Immediately after general relativity became known, *Willem de Sitter* (1872–1934) published (in 1917) another solution, for the case of empty space-time in an exponential state of expansion. We shall describe this solution in Section 4.2. In 1922 the Russian meteorologist *Alexandr Friedmann* (1888–1925) found a range of intermediate solutions to Einstein's equations which describe the standard cosmology today. Curiously, this work was ignored for a decade although it was published in widely read journals. This is the subject of Section 4.1.

In 1924 Hubble had measured the distances to nine spiral galaxies, and he found that they were extremely far away. The nearest one, M31 in the Andromeda, is now known to be at a distance of 20 galactic diameters (Hubble's value was about 8) and the farther ones at hundreds of galactic diameters. These observations established that the spiral nebulae are, as Kant had conjectured, stellar systems comparable in mass and size with the Milky Way, and their spatial distribution confirmed the expectations of the cosmological principle on the scale of galactic distances.

In 1926–1927 *Bertil Lindblad* (1895–1965) and *Jan Hendrik Oort* (1900–1992) verified Laplace's hypothesis that the Galaxy indeed rotated, and they determined the period to be 10^8 yr and the mass to be about $10^{11} M_\odot$. The conclusive demonstration that the Milky Way is an average-sized galaxy, in no way exceptional or central, was given only in 1952 by Walter Baade. This we may count as the third breakdown of the anthropocentric world picture.

The later history of cosmology up until 1990 has been excellently summarized by Peebles [3].

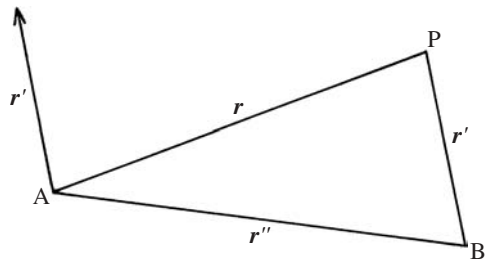


Figure 1.1 Two observers at A and B making observations in the directions $\boldsymbol{r}$, $\boldsymbol{r}'$.

To give the reader an idea of where in the Universe we are, what is nearby and what is far away, some cosmic distances are listed in Table A.1 in the appendix. On a cosmological scale we are not really interested in objects smaller than a galaxy! We generally measure cosmic distances in *parsec* (pc) units (kpc for 10^3 pc and Mpc for 10^6 pc). A parsec is the distance at which one second of arc is subtended by a length equalling the mean distance between the Sun and the Earth. The parsec unit is given in Table A.2 in the appendix, where the values of some useful cosmological and astrophysical constants are listed.

1.2 Inertial Frames and the Cosmological Principle

Newton's first law—the law of inertia—states that a system on which no forces act is either at rest or in uniform motion. Such systems are called *inertial frames*. Accelerated or rotating frames are not inertial frames. Newton considered that 'at rest' and 'in motion' implicitly referred to an *absolute space* which was unobservable but which had a real existence independent of humankind. Mach rejected the notion of an empty, unobservable space, and only Einstein was able to clarify the physics of motion of observers in inertial frames.

It may be interesting to follow a nonrelativistic argument about the static or nonstatic nature of the Universe which is a direct consequence of the cosmological principle.

Consider an observer 'A' in an inertial frame who measures the density of galaxies and their velocities in the space around him. Because the distribution of galaxies is observed to be homogeneous and isotropic on very large scales (strictly speaking, this is actually true for galaxy groups [1]), he would see the same mean density of galaxies (at one time t) in two different directions $\boldsymbol{r}$ and $\boldsymbol{r}'$:

$$\rho_{\mathrm{A}}(\boldsymbol{r}, t) = \rho_{\mathrm{A}}(\boldsymbol{r}', t).$$

Another observer 'B' in another inertial frame (see Figure 1.1) looking in the direction $\boldsymbol{r}$ from her location would also see the same mean density of galaxies:

$$\rho_{\mathrm{B}}(\boldsymbol{r}', t) = \rho_{\mathrm{A}}(\boldsymbol{r}, t).$$

The velocity distributions of galaxies would also look the same to both observers, in fact in all directions, for instance in the $\boldsymbol{r}'$ direction:

$$\boldsymbol{v}_{\mathrm{B}}(\boldsymbol{r}', t) = \boldsymbol{v}_{\mathrm{A}}(\boldsymbol{r}', t).$$

Suppose that the B frame has the relative velocity $\boldsymbol{v}_{\mathrm{A}}(\boldsymbol{r}'', t)$ as seen from the A frame along the radius vector $\boldsymbol{r}'' = \boldsymbol{r} - \boldsymbol{r}'$. If all velocities are nonrelativistic, i.e. small compared with the speed of light, we can write

$$\boldsymbol{v}_{\mathrm{A}}(\boldsymbol{r}', t) = \boldsymbol{v}_{\mathrm{A}}(\boldsymbol{r} - \boldsymbol{r}'', t) = \boldsymbol{v}_{\mathrm{A}}(\boldsymbol{r}, t) - \boldsymbol{v}_{\mathrm{A}}(\boldsymbol{r}'', t).$$

This equation is true only if $\boldsymbol{v}_{\mathrm{A}}(\boldsymbol{r}, t)$ has a specific form: it must be proportional to $\boldsymbol{r}$,

$$\boldsymbol{v}_{\mathrm{A}}(\boldsymbol{r}, t) = f(t)\boldsymbol{r}, \tag{1.1}$$

where $f(t)$ is an arbitrary function. Why is this so?

Let this universe start to expand. From the vantage point of A (or B equally well, since all points of observation are equal), nearby galaxies will appear to recede slowly. But in order to preserve uniformity, distant ones must recede faster, in fact their recession velocities must increase linearly with distance. That is the content of Equation (1.1).

If $f(t) > 0$, the Universe would be seen by both observers to expand, each galaxy having a radial velocity proportional to its radial distance $\boldsymbol{r}$. If $f(t) < 0$, the Universe would be seen to contract with velocities in the reversed direction. Thus we have seen that expansion and contraction are natural consequences of the cosmological principle. If $f(t)$ is a positive constant, Equation (1.1) is Hubble's law, which we shall meet in Section 1.4.

Actually, it is somewhat misleading to say that the galaxies recede when, rather, it is space itself which expands or contracts. This distinction is important when we come to general relativity.

A useful lesson may be learned from studying the limited gravitational system consisting of the Earth and rockets launched into space. This system is not quite like the previous example because it is not homogeneous, and because the motion of a rocket or a satellite in Earth's gravitational field is different from the motion of galaxies in the gravitational field of the Universe. Thus to simplify the case we only consider radial velocities, and we ignore Earth's rotation. Suppose the rockets have initial velocities low enough to make them fall back onto Earth. The rocket–Earth gravitational system is then *closed* and contracting, corresponding to $f(t) < 0$.

When the kinetic energy is large enough to balance gravity, our idealized rocket becomes a satellite, staying above Earth at a fixed height (real satellites circulate in stable Keplerian orbits at various altitudes if their launch velocities are in the range 8–11 km s^{-1}). This corresponds to the static solution $f(t) = 0$ for the rocket–Earth gravitational system.

If the launch velocities are increased beyond about 11 km s^{-1}, the potential energy of Earth's gravitational field no longer suffices to keep the rockets bound to Earth. Beyond this speed, called the *second cosmic velocity* by rocket engineers, the rockets escape for good. This is an expanding or *open* gravitational system, corresponding to $f(t) > 0$.

The static case is different if we consider the Universe as a whole. According to the cosmological principle, no point is preferred, and therefore there exists no centre around which bodies can gravitate in steady-state orbits. Thus the Universe

is either expanding or contracting, the static solution being unstable and therefore unlikely.

1.3 Olbers' Paradox

Let us turn to an early problem still discussed today, which is associated with the name of *Wilhelm Olbers* (1758–1840), although it seems to have been known already to Kepler in the 17th century, and a treatise on it was published by *Jean-Philippe Loys de Chéseaux* in 1744, as related in the book by E. Harrison [5]. Why is the night sky dark if the Universe is infinite, static and uniformly filled with stars? They should fill up the total field of visibility so that the night sky would be as bright as the Sun, and we would find ourselves in the middle of a heat bath of the temperature of the surface of the Sun. Obviously, at least one of the above assumptions about the Universe must be wrong.

The question of the total number of shining stars was already pondered by Newton and Leibnitz. Let us follow in some detail the argument published by Olbers in 1823. The *absolute luminosity* of a star is defined as the amount of luminous energy radiated per unit time, and the *surface brightness* B as luminosity per unit surface. Suppose that the number of stars with average luminosity L is N and their average density in a volume V is $n = N/V$. If the surface area of an average star is A, then its brightness is $B = L/A$. The Sun may be taken to be such an average star, mainly because we know it so well.

The number of stars in a spherical shell of radius r and thickness $\mathrm{d}r$ is then $4\pi r^2 n\,\mathrm{d}r$. Their total radiation as observed at the origin of a static universe of infinite extent is then found by integrating the spherical shells from 0 to ∞:

$$\int_0^\infty 4\pi r^2 nB\,\mathrm{d}r = \int_0^\infty nL\,\mathrm{d}r = \infty. \tag{1.2}$$

On the other hand, a finite number of visible stars each taking up an angle A/r^2 could cover an infinite number of more distant stars, so it is not correct to integrate r to ∞. Let us integrate only up to such a distance R that the whole sky of angle 4π would be evenly tiled by the star discs. The condition for this is

$$\int_0^R 4\pi r^2 n \frac{A}{r^2}\,\mathrm{d}r = 4\pi.$$

It then follows that the distance is $R = 1/An$. The integrated brightness from these visible stars alone is then

$$\int_0^R nL\,\mathrm{d}r = L/A, \tag{1.3}$$

or equal to the brightness of the Sun. But the night sky is indeed dark, so we are faced with a paradox.

Olbers' own explanation was that invisible interstellar dust absorbed the light. That would make the intensity of starlight decrease exponentially with distance. But one can show that the amount of dust needed would be so great that the Sun

would also be obscured. Moreover, the radiation would heat the dust so that it would start to glow soon enough, thereby becoming visible in the infrared.

A large number of different solutions to this paradox have been proposed in the past, some of the wrong ones lingering on into the present day. Let us here follow a valid line of reasoning due to Lord Kelvin (1824–1907), as retold and improved in a popular book by E. Harrison [5].

A star at distance r covers the fraction $A/4\pi r^2$ of the sky. Multiplying this by the number of stars in the shell, $4\pi r^2 n\,\mathrm{d}r$, we obtain the fraction of the whole sky covered by stars viewed by an observer at the centre, $An\,\mathrm{d}r$. Since n is the star count per volume element, An has the dimensions of number of stars per linear distance. The inverse of this,

$$\ell = 1/An, \tag{1.4}$$

is the mean radial distance between stars, or the *mean free path* of photons emitted from one star and being absorbed in collisions with another. We can also define a mean collision time:

$$\bar{\tau} = \ell/c. \tag{1.5}$$

The value of $\bar{\tau}$ can be roughly estimated from the properties of the Sun, with radius $R_\odot$ and density $\rho_\odot$. Let the present mean density of luminous matter in the Universe be ρ_0 and the distance to the farthest visible star r_*. Then the collision time inside this volume of size $\frac{4}{3}\pi r_*^3$ is

$$\bar{\tau} \simeq \bar{\tau}_\odot = \frac{1}{A_\odot n c} = \frac{1}{\pi R_\odot^2}\frac{4\pi r_*^3}{3Nc} = \frac{4\rho_\odot R_\odot}{3\rho_0 c}. \tag{1.6}$$

Taking the solar parameters from Table A.2 in the appendix we obtain approximately 10^{23} yr.

The probability that a photon does not collide but arrives safely to be observed by us after a flight distance r can be derived from the assumption that the photon encounters obstacles randomly, that the collisions occur independently and at a constant rate ℓ^{-1} per unit distance. The probability $P(r)$ that the distance to the first collision is r is then given by the exponential distribution

$$P(r) = \ell^{-1}\mathrm{e}^{-r/\ell}. \tag{1.7}$$

Thus flight distances much longer than ℓ are improbable.

Applying this to photons emitted in a spherical shell of thickness $\mathrm{d}r$, and integrating the spherical shell from zero radius to r_*, the fraction of all photons emitted in the direction of the centre of the sphere and arriving there to be detected is

$$f(r_*) = \int_0^{r_*} \ell^{-1}\mathrm{e}^{-r/\ell}\mathrm{d}r = 1 - \mathrm{e}^{-r_*/\ell}. \tag{1.8}$$

Obviously, this fraction approaches 1 only in the limit of an infinite universe. In that case every point on the sky would be seen to be emitting photons, and the sky would indeed be as bright as the Sun at night. But since this is not the case, we must conclude that r_*/ℓ is small. Thus the reason why the whole field of vision

is not filled with stars is that the volume of the presently observable Universe is not infinite, it is in fact too small to contain sufficiently many visible stars.

Lord Kelvin's original result follows in the limit of small r_*/ℓ, in which case

$$f(r_*) \approx r/\ell.$$

The exponential effect in Equation (1.8) was neglected by Lord Kelvin.

We can also replace the mean free path in Equation (1.8) with the collision time (1.5), and the distance r_* with the age of the Universe t_0, to obtain the fraction

$$f(r_*) = g(t_0) = 1 - \mathrm{e}^{-t_0/\bar{\tau}}. \tag{1.9}$$

If $u_\odot$ is the average radiation density at the surface of the stars, then the radiation density u_0 measured by us is correspondingly reduced by the fraction $g(t_0)$:

$$u_0 = u_\odot(1 - \mathrm{e}^{-t_0/\bar{\tau}}). \tag{1.10}$$

In order to be able to observe a luminous night sky we must have $u_0 \approx u_\odot$, or the Universe must have an age of the order of the collision time, $t_0 \approx 10^{23}$ yr. However, this exceeds all estimates of the age of the Universe (some estimates will be given in Section 1.5) by 13 orders of magnitude! Thus the existing stars have not had time to radiate long enough.

What Olbers and many after him did not take into account is that even if the age of the Universe was infinite, the stars do have a finite age and they burn their fuel at well-understood rates.

If we replace 'stars' by 'galaxies' in the above argument, the problem changes quantitatively but not qualitatively. The intergalactic space is filled with radiation from the galaxies, but there is less of it than one would expect for an infinite Universe, at all wavelengths. There is still a problem to be solved, but it is not quite as paradoxical as in Olbers' case.

One explanation is the one we have already met: each star radiates only for a finite time, and each galaxy has existed only for a finite time, whether the age of the Universe is infinite or not. Thus when the time perspective grows, an increasing number of stars become visible because their light has had time to reach us, but at the same time stars which have burned their fuel disappear.

Another possible explanation evokes expansion and special relativity. If the Universe expands, starlight redshifts, so that each arriving photon carries less energy than when it was emitted. At the same time, the volume of the Universe grows, and thus the energy density decreases. The observation of the low level of radiation in the intergalactic space has in fact been evoked as a proof of the expansion.

Since both explanations certainly contribute, it is necessary to carry out detailed quantitative calculations to establish which of them is more important. Most of the existing literature on the subject supports the relativistic effect, but Harrison has shown (and P. S. Wesson [6] has further emphasized) that this is false: the finite lifetime of the stars and galaxies is the dominating effect. The relativistic effect is quantitatively so unimportant that one cannot use it to prove that the Universe is either expanding or contracting.

1.4 Hubble's Law

In the 1920s Hubble measured the spectra of 18 spiral galaxies with a reasonably well-known distance. For each galaxy he could identify a known pattern of atomic spectral lines (from their relative intensities and spacings) which all exhibited a common redward frequency shift by a factor $1 + z$. Using the relation (1.1) following from the assumption of homogeneity alone,

$$v = cz, \tag{1.11}$$

he could then obtain their velocities with reasonable precision.

The Expanding Universe. The expectation for a stationary universe was that galaxies would be found to be moving about randomly. However, some observations had already shown that most galaxies were redshifted, thus receding, although some of the nearby ones exhibited blueshift. For instance, the nearby Andromeda nebula M31 is approaching us, as its blueshift testifies. Hubble's fundamental discovery was that the velocities of the distant galaxies he had studied increased linearly with distance:

$$v = H_0 r. \tag{1.12}$$

This is called *Hubble's law* and H_0 is called the *Hubble parameter*. For the relatively nearby spiral galaxies he studied, he could only determine the linear, first-order approximation to this function. Although the linearity of this law has been verified since then by the observations of hundreds of galaxies, it is not excluded that the true function has terms of higher order in r. In Section 2.3 we shall introduce a second-order correction.

The message of Hubble's law is that the Universe is expanding, and this general expansion is called the *Hubble flow*. At a scale of tens or hundreds of Mpc the distances to all astronomical objects are increasing regardless of the position of our observation point. It is true that we observe that the galaxies are receding *from us* as if we were at the centre of the Universe. However, we learned from studying a homogeneous and isotropic Universe in Figure 1.1 that if observer A sees the Universe expanding with the factor $f(t)$ in Equation (1.1), any other observer B will also see it expanding with the same factor, and the triangle ABP in Figure 1.1 will preserve its form. Thus, taking the cosmological principle to be valid, every observer will have the impression that all astronomical objects are receding from him/her. A homogeneous and isotropic Universe does not have a centre. Consequently, we shall usually talk about *expansion velocities* rather than *recession velocities*.

It is surprising that neither Newton nor later scientists, pondering about why the Universe avoided a gravitational collapse, came to realize the correct solution. An expanding universe would be slowed down by gravity, so the inevitable collapse would be postponed until later. It was probably the notion of an infinite scale of time, inherent in a stationary model, which blocked the way to the right conclusion.

Hubble Time and Radius. From Equations (1.11) and (1.12) one sees that the Hubble parameter has the dimension of inverse time. Thus a characteristic timescale for the expansion of the Universe is the *Hubble time*:

$$\tau_{\rm H} \equiv H_0^{-1} = 9.78h^{-1} \times 10^9 \text{ yr}. \tag{1.13}$$

Here h is the commonly used dimensionless quantity

$$h = H_0/(100 \text{ km s}^{-1} \text{ Mpc}^{-1}).$$

The Hubble parameter also determines the size scale of the observable Universe. In time $\tau_{\rm H}$, radiation travelling with the speed of light c has reached the *Hubble radius*:

$$r_{\rm H} \equiv \tau_{\rm H} c = 3000h^{-1} \text{ Mpc}. \tag{1.14}$$

Or, to put it a different way, according to Hubble's nonrelativistic law, objects at this distance would be expected to attain the speed of light, which is an absolute limit in the theory of special relativity.

Combining Equation (1.12) with Equation (1.11), one obtains

$$z = H_0 \frac{r}{c}. \tag{1.15}$$

In Section 2.1 on Special Relativity we will see limitations to this formula when v approaches c. The redshift z is in fact infinite for objects at distance $r_{\rm H}$ receding with the speed of light and thus physically meaningless. Therefore no information can reach us from farther away, all radiation is redshifted to infinite wavelengths, and no particle emitted within the Universe can exceed this distance.

The Cosmic Scale. The size of the Universe is unknown and unmeasurable, but if it undergoes expansion or contraction it is convenient to express distances at different epochs in terms of a *cosmic scale* $R(t)$, and denote its present value $R_0 \equiv R(t_0)$. The value of $R(t)$ can be chosen arbitrarily, so it is often more convenient to normalized it to its present value, and thereby define a dimensionless quantity, the *cosmic scale factor*:

$$a(t) \equiv R(t)/R_0. \tag{1.16}$$

The cosmic scale factor affects all distances: for instance the wavelength λ of light emitted at one time t and observed as λ_0 at another time t_0:

$$\frac{\lambda_0}{R_0} = \frac{\lambda}{R(t)}.$$

Let us find an approximation for $a(t)$ at times $t < t_0$ by expanding it to first-order time differences,

$$a(t) \approx 1 - \dot{a}_0(t_0 - t), \tag{1.17}$$

using the notation $\dot{a}_0$ for $\dot{a}(t_0)$, and $r = c(t_0 - t)$ for the distance to the source. The *cosmological redshift* can be approximated by

$$z = \frac{\lambda_0}{\lambda} - 1 = a^{-1} - 1 \approx \dot{a}_0 \frac{r}{c}. \tag{1.18}$$

Thus $1/1+z$ is a measure of the scale factor $a(t)$ at the time when a source emitted the now-redshifted radiation. Identifying the expressions for z in Equations (1.18) and (1.15) we find the important relation

$$\dot{a}_0 = \frac{\dot{R}_0}{R_0} = H_0. \tag{1.19}$$

The Hubble Constant. The value of this constant initially found by Hubble was $H_0 = 550\ \mathrm{km\ s^{-1}\ Mpc^{-1}}$: an order of magnitude too large because his distance measurements were badly wrong. To establish the linear law and to determine the global value of H_0 one needs to be able to measure distances and expansion velocities well and far out. Distances are precisely measured only to nearby stars which participate in the general rotation of the Galaxy, and which therefore do not tell us anything about cosmological expansion. Even at distances of several Mpc the expansion-independent, transversal *peculiar velocities* of galaxies are of the same magnitude as the Hubble flow. The measured expansion at the Virgo supercluster, 17 Mpc away, is about 1100 km s^{-1}, whereas the peculiar velocities attain 600 km s^{-1}. At much larger distances where the peculiar velocities do not contribute appreciably to the total velocity, for instance at the Coma cluster 100 Mpc away, the expansion velocity is 6900 km s^{-1} and the Hubble flow can be measured quite reliably, but the imprecision in distance measurements becomes the problem. Every procedure is sensitive to small, subtle corrections and to systematic biases unless great care is taken in the reduction and analysis of data.

A notable contribution to our knowledge of H_0 comes from the Hubble Space Telescope (HST) Key Project [7]. The goal of this project was to determine H_0 by a Cepheid calibration of a number of independent, secondary distance indicators, including Type Ia supernovae, the Tully–Fisher relation, the fundamental plane for elliptical galaxies, surface-brightness fluctuations, and Type-II supernovae. Here I shall restrict the discussion to the best absolute determinations of H_0, which are those from far away *supernovae*. (Cepheid distance measurements are discussed in Section 2.3 under the heading 'Distance Ladder Continued'.)

Occasionally, a very bright *supernova* explosion can be seen in some galaxy. These events are very brief (one month) and very rare: historical records show that in our Galaxy they have occurred only every 300 yr. The most recent nearby supernova occurred in 1987 (code name SN1987A), not exactly in our Galaxy but in our small satellite, the Large Magellanic Cloud (LMC). Since it has now become possible to observe supernovae in very distant galaxies, one does not have to wait 300 yr for the next one.

The physical reason for this type of explosion (a Type SNII supernova) is the accumulation of Fe group elements at the core of a massive red giant star of size 8–200$M_\odot$, which has already burned its hydrogen, helium and other light elements. Another type of explosion (a Type SNIa supernova) occurs in binary star systems, composed of a heavy white dwarf and a red giant star. White dwarfs have masses of the order of the Sun, but sizes of the order of Earth, whereas red

giants are very large but contain very little mass. The dwarf then accretes mass from the red giant due to its much stronger gravitational field.

As long as the fusion process in the dwarf continues to burn lighter elements to Fe group elements, first the gas pressure and subsequently the electron degeneracy pressure balance the gravitational attraction (degeneracy pressure is explained in Section 5.3). But when a rapidly burning dwarf star reaches a mass of $1.44M_\odot$, the so-called *Chandrasekhar mass*, or in the case of a red giant when the iron core reaches that mass, no force is sufficient to oppose the gravitational collapse. The electrons and protons in the core transform into neutrinos and neutrons, respectively, most of the gravitational energy escapes in the form of neutrinos, and the remainder is a *neutron star* which is stabilized against further gravitational collapse by the degeneracy pressure of the neutrons. As further matter falls in, it bounces against the extremely dense neutron star and travels outwards as energetic shock waves. In the collision between the shock waves and the outer mantle, violent nuclear reactions take place and extremely bright light is generated. This is the supernova explosion visible from very far away. The nuclear reactions in the mantle create all the elements; in particular, the elements heavier than Fe, Ni and Cr on Earth have all been created in supernova explosions in the distant past.

The released energy is always the same since the collapse always occurs at the Chandrasekhar mass, thus in particular the peak brightness of Type Ia supernovae can serve as remarkably precise standard candles visible from very far away. (The term *standard candle* is used for any class of astronomical objects whose intrinsic luminosity can be inferred independently of the observed flux.) Additional information is provided by the colour, the spectrum, and an empirical correlation observed between the timescale of the supernova light curve and the peak luminosity. The usefulness of supernovae of Type Ia as standard candles is that they can be seen out to great distances, 500 Mpc or $z \approx 0.1$, and that the internal precision of the method is very high. At greater distances one can still find supernovae, but Hubble's linear law (1.15) is no longer valid—the expansion starts to accelerate.

The SNeIa are the brightest and most homogeneous class of supernovae. (The plural of SN is abbreviated SNe.) Type II are fainter, and show a wider variation in luminosity. Thus they are not standard candles, but the time evolution of their expanding atmospheres provides an indirect distance indicator, useful out to some 200 Mpc.

Two further methods to determine H_0 make use of correlations between different galaxy properties. Spiral galaxies rotate, and there the *Tully–Fisher relation* correlates total luminosity with maximum rotation velocity. This is currently the most commonly applied distance indicator, useful for measuring extragalactic distances out to about 150 Mpc. Elliptical galaxies do not rotate, they are found to occupy a *fundamental plane* in which an effective radius is tightly correlated with the surface brightness inside that radius and with the central velocity dispersion of the stars. In principle, this method could be applied out to $z \approx 1$, but

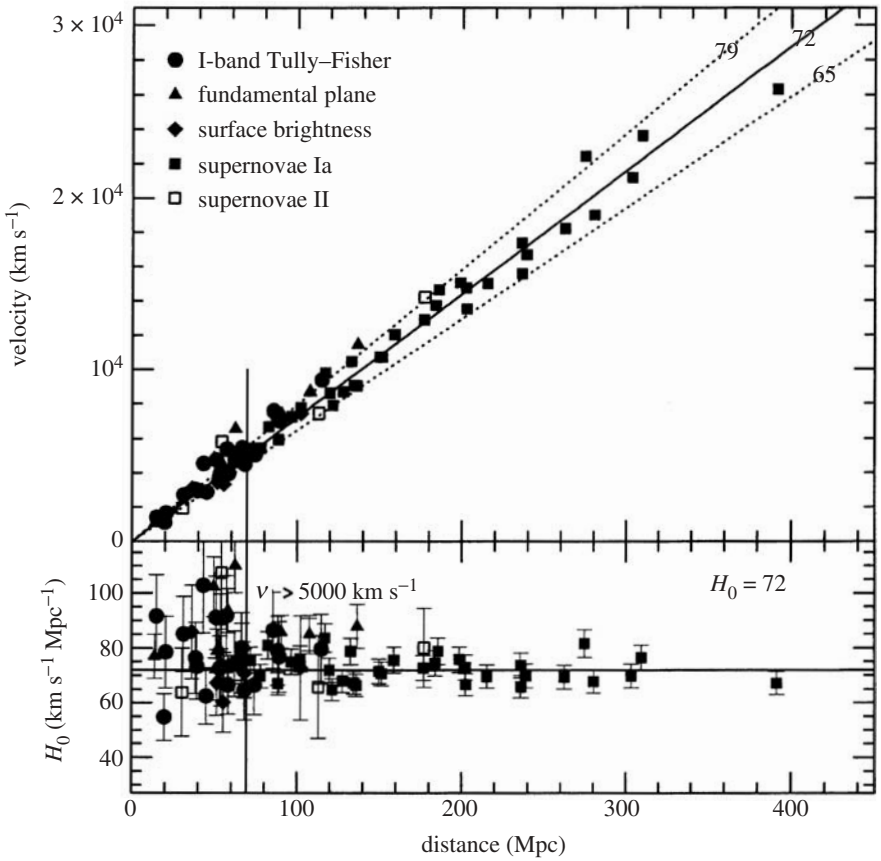


Figure 1.2 Recession velocities of different objects as a function of distance [7]. The slope determines the value of the Hubble constant.

in practice stellar evolution effects and the nonlinearity of Hubble's law limit the method to $z \lesssim 0.1$, or about 400 Mpc.

The resolution of individual stars within galaxies clearly depends on the distance to the galaxy. This method, called *surface-brightness fluctuations (SBFs)*, is an indicator of relative distances to elliptical galaxies and some spirals. The internal precision of the method is very high, but it can be applied only out to about 70 Mpc.

The observations of the HST have been confirmed by independent SNIa observations from observatories on the ground [8]. The HST team quotes

$$h \equiv H_0/(100 \text{ km s}^{-1} \text{ Mpc}^{-1}) = 0.72 \pm 0.03 \pm 0.07. \tag{1.20}$$

At the time of writing, even more precise determinations of H_0, albeit not significantly different, come from combined multiparameter analyses of the cosmic microwave background spectrum [9] and large-scale structures, to which we shall return in Chapters 8 and 9. The present best value, $h = 0.71$, is given in Equa-

tion (8.43) and in Table A.2 in the appendix. In Figure 1.2 we plot the combined HST observations of H_0.

Note that the second error in Equation (1.20), which is systematic, is much bigger than the statistical error. This illustrates that there are many unknown effects which complicate the determination of H_0, and which in the past have made all determinations controversial. To give just one example, if there is dust on the sight line to a supernova, its light would be reddened and one would conclude that the recession velocity is higher than it is in reality. There are other methods, such as weak lensing (to be discussed in Section 3.3), which do not suffer from this systematic error, but they have not yet reached a precision superior to that reported in Equation (1.20).

1.5 The Age of the Universe

One of the conclusions of Olbers' paradox was that the Universe could not be eternal, it must have an age much less than 10^{23} yr, or else the night sky would be bright. More recent proofs that the Universe indeed grows older and consequently has a finite lifetime comes from astronomical observations of many types of extragalactic objects at high redshifts and at different wavelengths: radio sources, X-ray sources, quasars, faint blue galaxies. High redshifts correspond to earlier times, and what are observed are clear changes in the populations and the characteristics as one looks toward earlier epochs. Let us therefore turn to determinations of the age of the Universe.

In Equation (1.13) we defined the Hubble time $\tau_{\rm H}$, and gave a value for it of the order of 10 billion years. However, $\tau_{\rm H}$ is not the same as the age t_0 of the Universe. The latter depends on the dynamics of the Universe, whether it is expanding forever or whether the expansion will turn into a collapse, and these scenarios depend on how much matter there is and what the geometry of the Universe is, all questions we shall come back to later. Taking h to be in the range 0.68–0.75, Equation (1.13) gives

$$t_0 \approx \tau_{\rm H} = 13.0\text{–}14.4 \text{ Gyr}. \tag{1.21}$$

Cosmochronology by Radioactive Nuclei. There are several independent techniques, *cosmochronometers*, for determining the age of the Universe. At this point we shall only describe determinations via the cosmochronology of long-lived radioactive nuclei, and via stellar modelling of the oldest stellar populations in our Galaxy and in some other galaxies. Note that the very existence of radioactive nuclides indicates that the Universe cannot be infinitely old and static.

Various nuclear processes have been used to date the age of the Galaxy, $t_{\rm G}$, for instance the 'Uranium clock'. Long-lived radioactive isotopes such as ^{232}Th, ^{235}U, ^{238}U and ^{244}Pu have been formed by fast neutrons from supernova explosions, captured in the envelopes of an early generation of stars. With each generation of star formation, burn-out and supernova explosion, the proportion of metals

increases. Therefore the metal-poorest stars found in globular clusters are the oldest.

The proportions of heavy isotopes following a supernova explosion are calculable with some degree of confidence. Since then, they have decayed with their different natural half-lives so that their abundances in the Galaxy today have changed. For instance, calculations of the original ratio $K = {}^{235}\mathrm{U}/{}^{238}\mathrm{U}$ give values of about 1.3 with a precision of about 10%, whereas this ratio on Earth at the present time is $K_0 = 0.007\,23$.

To compute the age of the Galaxy by this method, we also need the decay constants λ of ^{238}U and ^{235}U which are related to their half-lives:

$$\lambda_{238} = \ln 2/(4.46\ \mathrm{Gyr}), \qquad \lambda_{235} = \ln 2/(0.7038\ \mathrm{Gyr}).$$

The relation between isotope proportions, decay constants, and time t_G is

$$K = K_0 \exp[(\lambda_{238} - \lambda_{235})t_\mathrm{G}]. \tag{1.22}$$

Inserting numerical values one finds $t_\mathrm{G} \approx 6.2$ Gyr. However, the Solar System is only 4.57 Gyr old, so the abundance of ^{232}Th, ^{235}U and ^{238}U on Earth cannot be expected to furnish a very interesting limit to t_G. Rather, one has to turn to the abundances on the oldest stars in the Galaxy.

The globular clusters (GCs) are roughly spherically distributed stellar systems in the spheroid of the Galaxy. During the majority of the life of a star, it converts hydrogen into helium in its core. Thus the most interesting stars for the determination of t_G are those which have exhausted their supply of hydrogen, and which are located in old, metal-poor GCs, and to which the distance can be reliably determined. Over the last 10 yr, the GC ages have been reduced dramatically because of refined estimates of the parameters governing stellar evolution, and because of improved distance measurements. One can now quote [10] a best-fit age of 13.2 Gyr and a conservative lower limit of

$$t_\mathrm{GC} > 11.2\ \mathrm{Gyr}.$$

This includes an estimated age of greater than 0.8 Gyr for the Universe when the clusters formed.

Of particular interest is the detection of a spectral line of ^{238}U in the extremely metal-poor star CS 31082-001, which is overabundant in heavy elements [11]. Theoretical nucleosynthesis models for the initial abundances predict that the ratios of neighbouring stable and unstable elements should be similar in early stars as well as on Earth. Thus one compares the abundances of the radioactive ^{232}Th and ^{238}U with the neighbouring stable elements Os and Ir (^{235}U is now useless, because it has already decayed away on the oldest stars). One result [11] is that any age between 11.1 and 13.9 is compatible with the observations, whereas another group [12] using a different method quotes

$$t_* = 14.1 \pm 2.5\ \mathrm{Gyr}. \tag{1.23}$$

Bright Cluster Galaxies (BCGs). Another cosmochronometer is offered by the study of elliptical galaxies in BCGs at very large distances. It has been found that BCG colours only depend on their star-forming histories, and if one can trust stellar population synthesis models, one has a cosmochronometer. From an analysis of 17 bright clusters in the range $0.3 < z < 0.7$ observed by the HST, the result is [13]

$$t_{\mathrm{BCG}} = 13.4^{+1.4}_{-1.0}\ \mathrm{Gyr}. \tag{1.24}$$

Allowing 0.5–1 Gyr from the Big Bang until galaxies form stars and clusters, all the above three estimates fall in the range obtained from the Hubble constant in Equation (1.21). There are many more cosmochronometers making use of well-understood stellar populations at various distances which we shall not refer to here, all yielding ages near those quoted. It is of interest to note that in the past, when the dynamics of the Universe was less well known, the calculated age τ_{H} was smaller than the value in Equation (1.21), and at the same time the age t_* of the oldest stars was much higher than the value in Equation (1.23). Thus this historical conflict between cosmological and observational age estimates has now disappeared.

In Section 4.1 we will derive a general relativistic formula for t_0 which depends on a few measurable dynamical parameters. These parameters will only be defined later. They are determined in supernova analyses (in Section 4.4) and cosmic microwave background analyses (in Section 8.4). The best present estimate of t_0 is based on parameter values quoted by the Wilkinson Microwave Anisotropy Probe (WMAP) team [9]

$$t_0 = 13.7 \pm 0.2\ \mathrm{Gyr}. \tag{1.25}$$

1.6 Expansion in a Newtonian World

In this section we shall use Newtonian mechanics to derive a cosmology without recourse to Einstein's theory. Inversely, this formulation can also be derived from Einstein's theory in the limit of weak gravitational fields.

A system of massive bodies in an attractive Newtonian potential contracts rather than expands. The Solar System has contracted to a stable, gravitationally bound configuration from some form of hot gaseous cloud, and the same mechanism is likely to be true for larger systems such as the Milky Way, and perhaps also for clusters of galaxies. On yet larger scales the Universe expands, but this does not contradict Newton's law of gravitation.

The key question in cosmology is whether the Universe as a whole is a gravitationally bound system in which the expansion will be halted one day. We shall next derive a condition for this from Newtonian mechanics.

Newtonian Mechanics. Consider a galaxy of *gravitating mass* m_{G} located at a radius r from the centre of a sphere of mean density ρ and mass $M = 4\pi r^3\rho/3$

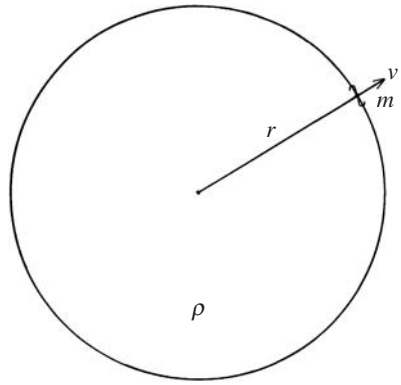


Figure 1.3 A galaxy of mass m at radial distance r receding with velocity v from the centre of a homogeneous mass distribution of density ρ.

(see Figure 1.3). The gravitational potential of the galaxy is

$$U = -GMm_{\mathrm{G}}/r = -\tfrac{4}{3}\pi G m_{\mathrm{G}}\rho r^2, \tag{1.26}$$

where G is the *Newtonian constant* expressing the strength of the gravitational interaction. Thus the galaxy falls towards the centre of gravitation, acquiring a radial acceleration

$$\ddot{r} = -GM/r^2 = -\tfrac{4}{3}\pi G\rho r. \tag{1.27}$$

This is *Newton's law of gravitation*, usually written in the form

$$F = -\frac{GMm_{\mathrm{G}}}{r^2}, \tag{1.28}$$

where F (in old-fashioned parlance) is the force exerted by the mass M on the mass m_{G}. The negative signs in Equations (1.26)–(1.28) express the attractive nature of gravitation: bodies are forced to move in the direction of decreasing r.

In a universe expanding linearly according to Hubble's law (Equation (1.12)), the kinetic energy T of the galaxy receding with velocity v is

$$T = \tfrac{1}{2}mv^2 = \tfrac{1}{2}mH_0^2r^2, \tag{1.29}$$

where m is the *inertial mass* of the galaxy. Although there is no theoretical reason for the inertial mass to equal the gravitational mass (we shall come back to this question later), careful tests have verified the equality to a precision better than a few parts in 10^{13}. Let us therefore set $m_{\mathrm{G}} = m$. Thus the total energy is given by

$$E = T + U = \tfrac{1}{2}mH_0^2r^2 - \tfrac{4}{3}\pi Gm\rho r^2 = mr^2(\tfrac{1}{2}H_0^2 - \tfrac{4}{3}\pi G\rho). \tag{1.30}$$

If the mass density ρ of the Universe is large enough, the expansion will halt. The condition for this to occur is $E = 0$, or from Equation (1.30) this *critical density* is

$$\rho_{\mathrm{c}} = \frac{3H_0^2}{8\pi G} = 1.0539 \times 10^{10}h^2 \text{ eV m}^{-3}. \tag{1.31}$$

A universe with density $\rho > \rho_{\mathrm{c}}$ is called *closed*; with density $\rho < \rho_{\mathrm{c}}$ it is called *open*.

Expansion. Note that r and ρ are time dependent: they scale with the expansion. Denoting their present values r_0 and ρ_0, one has

$$r(t) = r_0 a(t), \qquad \rho(t) = \rho_0 a^{-3}(t). \tag{1.32}$$

The acceleration $\ddot{r}$ in Equation (1.27) can then be replaced by the acceleration of the scale:

$$\ddot{a} = \ddot{r}/r_0 = -\tfrac{4}{3}\pi G a^{-2}. \tag{1.33}$$

Let us use the identity

$$\ddot{a} = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}a}\dot{a}^2$$

in Equation (1.33) to obtain

$$\mathrm{d}\dot{a}^2 = -\tfrac{8}{3}\pi G\rho_0 \frac{\mathrm{d}a}{a^2}.$$

This can be integrated from the present time t_0 to an earlier time t with the result

$$\dot{a}^2(t) - \dot{a}^2(t_0) = \tfrac{8}{3}\pi G\rho_0(a^{-1} - 1). \tag{1.34}$$

Let us now introduce the dimensionless *density parameter*:

$$\Omega_0 = \frac{\rho_0}{\rho_c} = \frac{8\pi G\rho_0}{3H_0^2}. \tag{1.35}$$

Substituting Ω_0 into Equation (1.34) and making use of the relation (1.19), $\dot{a}(t_0) = H_0$, we find

$$\dot{a}^2 = H_0^2(\Omega_0 a^{-1} - \Omega_0 + 1). \tag{1.36}$$

Thus it is clear that the presence of matter influences the dynamics of the Universe. Without matter, $\Omega_0 = 0$, Equation (1.36) just states that the expansion is constant, $\dot{a} = H_0$, and H_0 could well be zero as Einstein thought. During expansion $\dot{a}$ is positive; during contraction it is negative. In both cases the value of $\dot{a}^2$ is nonnegative, so it must always be true that

$$1 - \Omega_0 + \Omega_0/a \geqslant 0. \tag{1.37}$$

Cosmological Models. Depending on the value of Ω_0 the evolution of the Universe can take three courses.

(i) $\Omega_0 < 1$, the mass density is undercritical. As the cosmic scale factor $a(t)$ increases for times $t > t_0$ the term Ω_0/a decreases, but the expression (1.37) stays positive always. Thus this case corresponds to an open, ever-expanding universe, as a consequence of the fact that it is expanding now. In Figure 1.4 the expression (1.37) is plotted against a as the long-dashed curve for the choice $\Omega_0 = 0.5$.

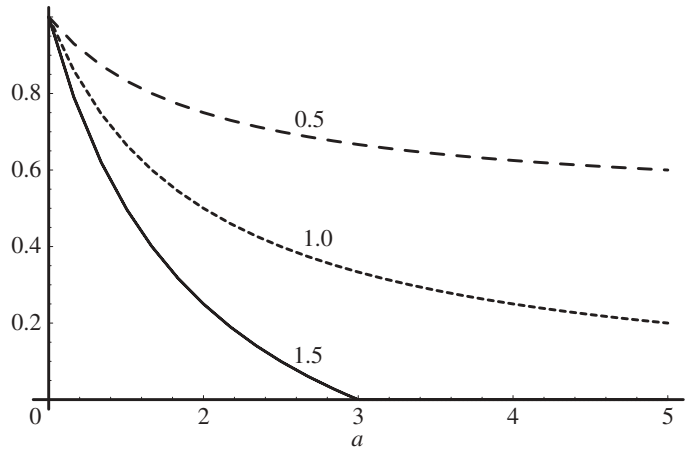


Figure 1.4 Dependence of the expression (1.37) on the cosmic scale a for an undercritical ($\Omega_0 = 0.5$), critical ($\Omega_0 = 1$) and overcritical ($\Omega_0 = 1.5$) universe. Time starts today at scale $a = 1$ in this picture and increases with a, except for the overcritical case where the Universe arrives at its maximum size, here $a = 3$, whereupon it reverses its direction and starts to shrink.

(ii) $\Omega_0 = 1$, the mass density is critical. As the scale factor $a(t)$ increases for times $t > t_0$ the expression in Equation (1.37) gradually approaches zero, and the expansion halts. However, this only occurs infinitely late, so it also corresponds to an ever-expanding universe. This case is plotted against a as the short-dashed curve in Figure 1.4. Note that cases (i) and (ii) differ by having different asymptotes. Case (ii) is quite realistic because the observational value of Ω_0 is very close to 1, as we shall see later.

(iii) $\Omega_0 > 1$, the mass density is overcritical and the Universe is closed. As the scale factor $a(t)$ increases, it reaches a maximum value $a_{\rm mid}$ when the expression in Equation (1.37) vanishes, and where the rate of increase, $\dot{a}_{\rm mid}$, also vanishes. But the condition (1.37) must stay true, and therefore the expansion must turn into contraction at $a_{\rm mid}$. The solid line in Figure 1.4 describes this case for the choice $\Omega_0 = 1.5$, whence $a_{\rm mid} = 3$. For later times the Universe retraces the solid curve, ultimately reaching scale $a = 1$ again.

This is as far as we can go combining Newtonian mechanics with Hubble's law. We have seen that problems appear when the recession velocities exceed the speed of light, conflicting with special relativity. Another problem is that Newton's law of gravitation knows no delays: the gravitational potential is felt instantaneously over all distances. A third problem with Newtonian mechanics is that the Copernican world, which is assumed to be homogeneous and isotropic, extends up to a finite distance r_0, but outside that boundary there is nothing. Then the boundary region is characterized by violent inhomogeneity and anisotropy, which are not taken into account. To cope with these problems we must begin to construct a fully relativistic cosmology.

Problems

1. How many revolutions has the Galaxy made since the formation of the Solar System if we take the solar velocity around the galactic centre to be 365 km s^{-1}?
2. Use Equation (1.4) to estimate the mean free path ℓ of photons. What fraction of all photons emitted by stars up to the maximum observed redshift $z = 7$ arrive at Earth?
3. If Hubble had been right that the expansion is given by

$$H_0 = 550 \text{ km s}^{-1}\text{Mpc}^{-1},$$

   how old would the Universe be then (see (1.13))?
4. What is the present ratio $K_0 = {}^{235}\text{U}/{}^{238}\text{U}$ on a star 10 Gyr old?
5. Prove Newton's theorem that the gravitational force at a radial distance R from the centre of a spherical distribution of matter acts as if all the mass inside R were concentrated at a single point at the centre. Show also that if the spherical distribution of matter extends beyond R, the force due to the mass outside R vanishes [14].
6. Estimate the escape velocity from the Galaxy.

Chapter Bibliography

[1] Ramella, M., Geller, M. J., Pisani, A. and da Costa, L. N. 2002 *Astron. J.* **123**, 2976.
[2] Fang Li Zhi and Li Shu Xian 1989 *Creation of the Universe*. World Scientific, Singapore.
[3] Peebles, P. J. E. 1993 *Principles of physical cosmology*. Princeton University Press, Princeton, NJ.
[4] Hagiwara, K. *et al*. 2002 *Phys. Rev.* D **66**, 010001-1.
[5] Harrison, E. 1987 *Darkness at night*. Harvard University Press, Cambridge, MA.
[6] Wesson, P. S. 1991 *Astrophys. J.* **367**, 399.
[7] Freedman, W. L. *et al*. 2001 *Astrophys. J.* **553**, 47.
[8] Gibson, B. K. and Brook, C. B. 2001 *New cosmological data and the values of the fundamental parameters* (ed. A. Lasenby & A. Wilkinson), ASP Conference Proceedings Series, vol. 666.
[9] Bennett, C. L. *et al*. 2003 Preprint arXiv, astro-ph/0302207 and 2003 *Astrophys. J.* (In press.) and companion papers cited therein.
[10] Krauss, L. M. and Chaboyer, B. 2003 *Science* **299**, 65–69.
[11] Cayrel, R. *et al*. 2001 *Nature* **409**, 691.
[12] Wanajo, S. *et al*. 2002 *Astrophys. J.* **577**, 853.
[13] Ferreras, I. *et al*. 2001 *Mon. Not. R. Astron. Soc.* **327**, L47.
[14] Shu, F. H. 1982 *The physical Universe*. University Science Books, Mill Valley, CA.

2
Relativity

The foundations of modern cosmology were laid during the second and third decade of the 20th century: on the theoretical side by Einstein's theory of general relativity, which represented a deep revision of current concepts; and on the observational side by Hubble's discovery of the cosmic expansion, which ruled out a static Universe and set the primary requirement on theory. Space and time are not invariants under Lorentz transformations, their values being different to observers in different inertial frames. Non-relativistic physics uses these quantities as completely adequate approximations, but in relativistic frame-independent physics we must find invariants to replace them. This chapter begins, in Section 2.1, with Einstein's theory of special relativity, which gives us such invariants.

In Section 2.2 we generalize the metrics in linear spaces to metrics in curved spaces, in particular the Robertson–Walker metric in a four-dimensional manifold. This gives us tools to define invariant distance measures in Section 2.3, and to conclude with a brief review of astronomical distance measurements which are the key to Hubble's parameter.

A central task of this chapter is to derive Einstein's law of gravitation using as few mathematical tools as possible (for far more detail, see, for example, [1] and [2]). The basic principle of covariance introduced in Section 2.4 requires a brief review of tensor analysis. Tensor notation has the advantage of permitting one to write laws of nature in the same form in all invariant systems.

The 'principle of equivalence' is introduced in Section 2.5 and it is illustrated by examples of travels in lifts. In Section 2.6 we assemble all these tools and arrive at Einstein's law of gravitation.

2.1 Lorentz Transformations and Special Relativity

In Einstein's theory of special relativity one studies how signals are exchanged between inertial frames in motion with respect to each other with constant velocity. Einstein made two postulates about such frames:

Introduction to Cosmology Third Edition by Matts Roos
© 2003 John Wiley & Sons, Ltd ISBN 0 470 84909 6 (cased) ISBN 0 470 84910 X (pbk)

(i) the results of measurements in different frames must be identical; and

(ii) light travels by a constant speed, c, in vacuo, in all frames.

The first postulate requires that physics be expressed in frame-independent invariants. The latter is actually a statement about the measurement of time in different frames, as we shall see shortly.

Lorentz Transformations. Consider two linear axes x and x' in one-dimensional space, x' being at rest and x moving with constant velocity v in the positive x' direction. Time increments are measured in the two coordinate systems as $\mathrm{d}t$ and $\mathrm{d}t'$ using two identical clocks. Neither the spatial increments $\mathrm{d}x$ and $\mathrm{d}x'$ nor the time increments $\mathrm{d}t$ and $\mathrm{d}t'$ are invariants—they do not obey postulate (i). Let us replace $\mathrm{d}t$ and $\mathrm{d}t'$ with the temporal distances $c\,\mathrm{d}t$ and $c\,\mathrm{d}t'$ and look for a *linear transformation* between the primed and unprimed coordinate systems, under which the two-dimensional *space-time distance* $\mathrm{d}\boldsymbol{s}$ between two *events*,

$$\mathrm{d}s^2 = c^2\,\mathrm{d}\tau^2 = c^2\,\mathrm{d}t^2 - \mathrm{d}x^2 = c^2\,\mathrm{d}t'^2 - \mathrm{d}x'^2, \tag{2.1}$$

is invariant. Invoking the constancy of the speed of light it is easy to show that the transformation must be of the form

$$\mathrm{d}x' = \gamma(\mathrm{d}x - v\,\mathrm{d}t), \qquad c\,\mathrm{d}t' = \gamma(c\,\mathrm{d}t - v\,\mathrm{d}x/c), \tag{2.2}$$

where

$$\gamma = \frac{1}{\sqrt{1-(v/c)^2}}. \tag{2.3}$$

Equation (2.2) defines the *Lorentz transformation*, after *Hendrik Antoon Lorentz* (1853–1928). Scalar products in this two-dimensional (ct, x)-space are invariants under Lorentz transformations.

Time Dilation. The quantity $\mathrm{d}\tau$ in Equation (2.1) is called the *proper time* and $\mathrm{d}\boldsymbol{s}$ the *line element*. Note that scalar multiplication in this manifold is here defined in such a way that the products of the spatial components obtain negative signs (sometimes the opposite convention is chosen). (The mathematical term for a many-dimensional space is a *manifold*.)

Since $\mathrm{d}\tau^2$ is an invariant, it has the same value in both frames:

$$\mathrm{d}\tau'^2 = \mathrm{d}\tau^2.$$

While the observer at rest records consecutive ticks on his clock separated by a space-time interval $\mathrm{d}\tau = \mathrm{d}t'$, she receives clock ticks from the x direction separated by the time interval $\mathrm{d}t$ and also by the space interval $\mathrm{d}x = v\,\mathrm{d}t$:

$$\mathrm{d}\tau = \mathrm{d}\tau' = \sqrt{\mathrm{d}t^2 - \mathrm{d}x^2/c^2} = \sqrt{1-(v/c)^2}\,\mathrm{d}t. \tag{2.4}$$

In other words, the two inertial coordinate systems are related by a Lorentz transformation

$$\mathrm{d}t = \frac{\mathrm{d}t'}{\sqrt{1-(v/c)^2}} \equiv \gamma\,\mathrm{d}t'.$$

Obviously, the time interval dt is always longer than the interval dt', but only noticeably so when v approaches c. This is called the *time dilation effect.*

The time dilation effect has been well confirmed in particle experiments. Muons are heavy, unstable, electron-like particles with well-known lifetimes in the laboratory. However, when they strike Earth with relativistic velocities after having been produced in cosmic ray collisions in the upper atmosphere, they appear to have a longer lifetime by the factor γ.

Another example is furnished by particles of mass m and charge Q circulating with velocity v in a synchrotron of radius r. In order to balance the centrifugal force the particles have to be subject to an inward-bending magnetic field density B. The classical condition for this is

$$r = mv/QB.$$

The velocity in the circular synchrotron as measured by a physicist at rest in the laboratory frame is inversely proportional to t, say the time of one revolution. But in the particle rest frame the time of one revolution is shortened to t/γ. When the particle attains relativistic velocities (by traversing accelerating potentials at regular positions in the ring), the magnetic field density B felt by the particle has to be adjusted to match the velocity in the particle frame, thus

$$r = mv\gamma/QB.$$

This equation has often been misunderstood to imply that the mass m increases by the factor γ, whereas only time measurements are affected by γ.

Relativity and Gold. Another example of relativistic effects on the orbits of circulating massive particles is furnished by electrons in Bohr orbits around a heavy nucleus. The effective Bohr radius of an electron is inversely proportional to its mass. Near the nucleus the electrons attain relativistic speeds, the time dilation will cause an apparent increase in the electron mass, more so for inner electrons with larger average speeds. For a 1s shell at the nonrelativistic limit, this average speed is proportional to Z atomic units. For instance, v/c for the 1s electron in Hg is $80/137 = 0.58$, implying a relativistic radial shrinkage of 23%. Because the higher s shells have to be orthogonal against the lower ones, they will suffer a similar contraction. Due to interacting relativistic and shell-structure effects, their contraction can be even larger; for gold, the 6s shell has larger percentage relativistic effects than the 1s shell. The nonrelativistic 5d and 6s orbital energies of gold are similar to the 4d and 5s orbital energies of silver, but the relativistic energies happen to be very different. This is the cause of the chemical difference between silver and gold and also the cause for the distinctive colour of gold [3].

Light Cone. The Lorentz transformations (2.1), (2.2) can immediately be generalized to three spatial dimensions, where the square of the Pythagorean distance element

$$dl^2 \equiv d\boldsymbol{l}^2 = dx^2 + dy^2 + dz^2 \tag{2.5}$$

is invariant under rotations and translations in three-space. This is replaced by the four-dimensional space-time of *Hermann Minkowski* (1864–1909), defined by the temporal distance ct and the spatial coordinates x, y, z. An invariant under Lorentz transformations between frames which are rotated or translated at a constant velocity with respect to each other is then the line element of the *Minkowski metric*:

$$\mathrm{d}s^2 = c^2\,\mathrm{d}\tau^2 = c^2\,\mathrm{d}t^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2 = c^2\,\mathrm{d}t^2 - \mathrm{d}l^2. \tag{2.6}$$

The trajectory of a body moving in space-time is called its *world line*. A body at a fixed location in space follows a world line parallel to the time axis and, of course, in the direction of increasing time. A body moving in space follows a world line making a slope with respect to the time axis. Since the speed of a body or a signal travelling from one event to another cannot exceed the speed of light, there is a maximum slope to such world lines. All world lines arriving where we are, here and now, obey this condition. Thus they form a cone in our past, and the envelope of the cone corresponds to signals travelling with the speed of light. This is called the *light cone*.

Two separate events in space-time can be *causally* connected provided their spatial separation $\mathrm{d}\boldsymbol{l}$ and their temporal separation $\mathrm{d}t$ (in any frame) obey

$$|\mathrm{d}\boldsymbol{l}/\mathrm{d}t| \leqslant c.$$

Their world line is then inside the light cone. In Figure 2.1 we draw this four-dimensional cone in t, x, y-space, but another choice would have been to use the coordinates t, σ, θ. Thus if we locate the *present* event at the apex of the light cone at $t = t_0 = 0$, it can be influenced by world lines from all events inside the *past* light cone for which $ct < 0$, and it can influence all events inside the *future* light cone for which $ct > 0$. Events inside the light cone are said to have *timelike* separation from the present event. Events outside the light cone are said to have *spacelike* separation from the present event: they cannot be causally connected to it. Thus the light cone encloses the *present observable universe*, which consists of all world lines that can in principle be observed. From now on we usually mean the present observable universe when we say simply 'the Universe'.

For light signals the equality sign above applies so that the proper time interval in Equation (2.6) vanishes:

$$\mathrm{d}\tau = 0.$$

Events on the light cone are said to have *null* or *lightlike* separation.

Redshift and Scale Factor. The light emitted by stars is caused by atomic transitions with emission spectra containing sharp spectral lines. Similarly, hot radiation traversing cooler matter in stellar atmospheres excites atoms at sharply defined wavelengths, producing characteristic dark absorption lines in the continuous regions of the emission spectrum. The radiation that was emitted by stars and distant galaxies with a wavelength $\lambda_{\mathrm{rest}} = c/\nu_{\mathrm{rest}}$ at time t in their rest frame will have its wavelength stretched by the cosmological expansion to λ_{obs} when

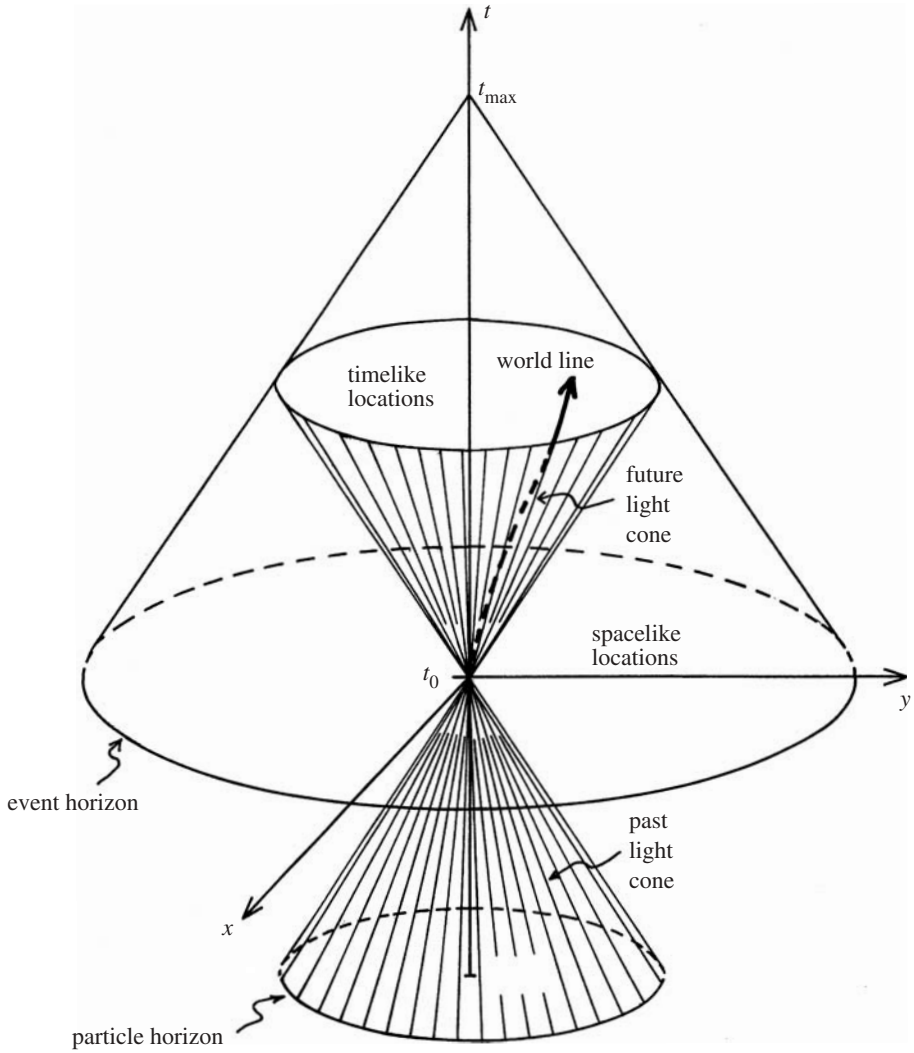


Figure 2.1 Light cone in x, y, t-space. An event which is at the origin $x = y = 0$ at the present time t_0 will follow some world line into the future, always remaining inside the future light cone. All points on the world line are at timelike locations with respect to the spatial origin at t_0. World lines for light signals emitted from (received at) the origin at t_0 will propagate on the envelope of the future (past) light cone. No signals can be sent to or received from spacelike locations. The space in the past from which signals can be received at the present origin is restricted by the particle horizon at t_{min}, the earliest time under consideration. The event horizon restricts the space which can at present be in causal relation to the present spatial origin at some future time t_{max}.

observed on Earth. Since the Universe expands, this shift is in the red direction, $\lambda_{obs} > \lambda_{rest}$, and it is therefore called a *redshift*, denoted

$$z = \frac{\lambda_{obs} - \lambda_{rest}}{\lambda_{rest}}. \tag{2.7}$$

The letter z for the redshift is of course a different quantity than the coordinate z in Equations (2.5) and (2.6).

The ratio of wavelengths actually measured by the terrestrial observer is then

$$1 + z = \frac{\lambda_{\text{obs}}}{\lambda_{\text{rest}}} = \frac{R_0}{R(t)} = \frac{1}{a(t)}. \tag{2.8}$$

It should be stressed that the cosmological redshift is not caused by the velocities of receding objects, only by the increase in scale $a(t)$ since time t. A kinematic effect can be observed in the spectra of nearby stars and galaxies, for which their peculiar motion is more important than the effect of the cosmological expansion. This may give rise to a *Doppler redshift* for a receding source, and to a corresponding *blueshift* for an approaching source.

Actually, the light cones in Figure 2.1 need to be modified for an expanding universe. A scale factor $a(t)$ that increases with time implies that light will travel a distance greater than ct during time t. Consequently, the straight lines will be curved.

2.2 Metrics of Curved Space-time

In Newton's time the laws of physics were considered to operate in a *flat Euclidean space*, in which spatial distance could be measured on an infinite and immovable three-dimensional grid, and time was a parameter marked out on a linear scale running from infinite past to infinite future. But Newton could not answer the question of how to identify which inertial frame was at rest relative to this absolute space. In his days the solar frame could have been chosen, but today we know that the Solar System orbits the Galactic centre, the Galaxy is in motion relative to the local galaxy group, which in turn is in motion relative to the Hydra–Centaurus cluster, and the whole Universe is expanding.

The geometry of curved spaces was studied in the 19th century by Gauss, Riemann and others. Riemann realized that Euclidean geometry was just a particular choice suited to flat space, but not necessarily correct in the space we inhabit. And Mach realized that one had to abandon the concept of absolute space altogether. Einstein learned about *tensors* from his friend Marcel Grossman, and used these key quantities to go from flat Euclidean three-dimensional space to curved Minkowskian four-dimensional space in which physical quantities are described by invariants. Tensors are quantities which provide generally valid relations between different four-vectors.

Euclidean Space. Let us consider how to describe distance in three-space. The path followed by a free body obeying Newton's first law of motion can suitably be described by expressing its spatial coordinates as functions of time: $x(t)$, $y(t)$, $z(t)$. Time is then treated as an absolute parameter and not as a coordinate. This path represents the shortest distance between any two points along it, and it is called a *geodesic* of the space. As is well known, in Euclidean space

the geodesics are straight lines. Note that the definition of a geodesic does not involve any particular coordinate system.

If one replaces the components x, y, z of the distance vector $\boldsymbol{l}$ by x^1, x^2, x^3, this permits a more compact notation of the Pythagorean squared distance l^2 in the *metric equation* (2.5):

$$l^2 = (x^1)^2 + (x^2)^2 + (x^3)^2 = \sum_{i,j=1}^{3} g_{ij} x^i x^j \equiv g_{ij} x^i x^j. \tag{2.9}$$

The quantities g_{ij} are the nine components of the *metric tensor* $\boldsymbol{g}$, which contains all the information about the intrinsic geometry of this three-space. In the last step we have used the convention to leave out the summation sign; it is then implied that summation is carried out over repeated indices. One commonly uses Roman letters in the indices when only the spatial components x^i, $i = 1, 2, 3$, are implied, and Greek letters when all the four space-time coordinates x^μ, $\mu = 0, 1, 2, 3$, are implied. Orthogonal coordinate systems have diagonal metric tensors and this is all that we will encounter. The components of $\boldsymbol{g}$ in flat Euclidean three-space are

$$g_{ij} = \delta_{ij},$$

where δ_{ij} is the usual Kronecker delta.

The same flat space could equally well be mapped by, for example, spherical or cylindrical coordinates. The components g_{ij} of the metric tensor would be different, but Equation (2.9) would hold unchanged. For instance, choosing spherical coordinates R, θ, ϕ as in Figure 2.2,

$$x = R \sin\theta \sin\phi, \qquad y = R \sin\theta \cos\phi, \qquad z = R \cos\theta, \tag{2.10}$$

$\mathrm{d}l^2$ takes the explicit form

$$\mathrm{d}l^2 = \mathrm{d}R^2 + R^2\,\mathrm{d}\theta^2 + R^2 \sin^2\theta\,\mathrm{d}\phi^2. \tag{2.11}$$

Geodesics in this space obey Newton's first law of motion, which may be written as

$$\ddot{\boldsymbol{R}} = 0. \tag{2.12}$$

Minkowski Space-Time. In special relativity, symmetry between spatial coordinates and time is achieved, as is evident from the Minkowski metric (2.6) describing a flat space-time in four Cartesian coordinates. In tensor notation the Minkowski metric includes the coordinate $\mathrm{d}x^0 \equiv c\,\mathrm{d}t$ so that the invariant line element in Equation (2.6) can be written

$$\mathrm{d}s^2 = c^2\,\mathrm{d}\tau^2 = g_{\mu\nu}\,\mathrm{d}x^\mu\,\mathrm{d}x^\nu. \tag{2.13}$$

The components of $\boldsymbol{g}$ in flat Minkowski space-time are given by the diagonal matrix $\eta_{\mu\nu}$, a generalization of the Kronecker delta function to four-space-time,

$$\eta_{00} = 1, \qquad \eta_{jj} = -1, \quad j = 1, 2, 3, \tag{2.14}$$

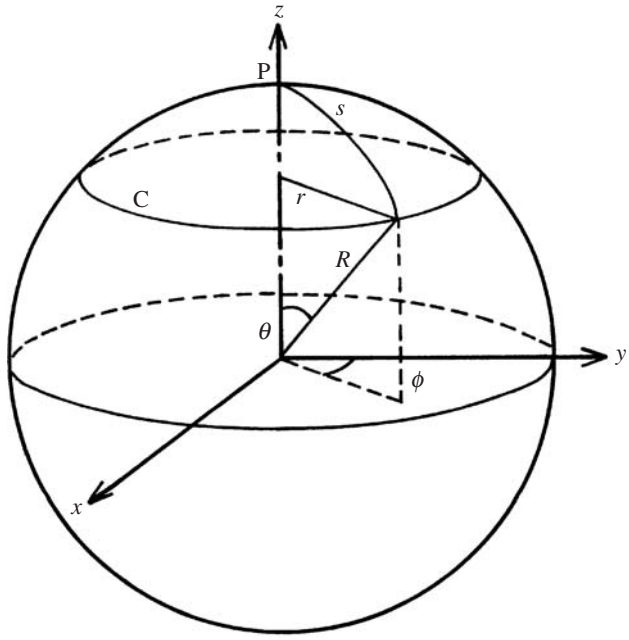


Figure 2.2 A two-sphere on which points are specified by coordinates (θ, ϕ).

all nondiagonal components vanishing. The choice of signs in the definition of $\eta_{\mu\nu}$ is not standardized in the literature, but we shall use Equation (2.14).

The path of a body, or its world line, is then described by the four coordinate functions $x(\tau)$, $y(\tau)$, $z(\tau)$, $t(\tau)$, where the proper time τ is a new absolute parameter, an invariant under Lorentz transformations. A geodesic in the Minkowski space-time is also a straight line, given by the equations

$$\frac{\mathrm{d}^2 t}{\mathrm{d}\tau^2} = 0, \qquad \frac{\mathrm{d}^2 \boldsymbol{R}}{\mathrm{d}\tau^2} = 0. \tag{2.15}$$

In the spherical coordinates (2.10) the Minkowski metric (2.6) takes the form

$$\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 - \mathrm{d}l^2 = c^2\,\mathrm{d}t^2 - \mathrm{d}R^2 - R^2\,\mathrm{d}\theta^2 - R^2 \sin^2\theta\,\mathrm{d}\phi^2. \tag{2.16}$$

An example of a curved space is the two-dimensional surface of a sphere with radius R obeying the equation

$$x^2 + y^2 + z^2 = R^2. \tag{2.17}$$

This surface is called a two-sphere.

Combining Equations (2.5) and (2.17) we see that one coordinate is really superfluous, for instance z, so that the spatial metric (2.5) can be written

$$\mathrm{d}l^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \frac{(x\,\mathrm{d}x + y\,\mathrm{d}y)^2}{R^2 - x^2 - y^2}. \tag{2.18}$$

This metric describes spatial distances on a two-dimensional surface embedded in three-space, but the third dimension is not needed to measure a distance on

the surface. Note that R is not a third coordinate, but a constant everywhere on the surface.

Thus measurements of distances depend on the geometric properties of space, as has been known to navigators ever since Earth was understood to be spherical. The geodesics on a sphere are great circles, and the metric is

$$\mathrm{d}l^2 = R^2\,\mathrm{d}\theta^2 + R^2\sin^2\theta\,\mathrm{d}\phi^2. \tag{2.19}$$

Near the poles where $\theta = 0^\circ$ or $\theta = 180^\circ$ the local distance would depend very little on changes in longitude ϕ. No point on this surface is preferred, so it can correspond to a Copernican homogeneous and isotropic two-dimensional universe which is unbounded, yet finite.

Let us write Equation (2.19) in the matrix form

$$\mathrm{d}l^2 = \begin{pmatrix}\mathrm{d}\theta & \mathrm{d}\phi\end{pmatrix} \boldsymbol{g} \begin{pmatrix}\mathrm{d}\theta \\ \mathrm{d}\phi\end{pmatrix}, \tag{2.20}$$

where the metric matrix is

$$\boldsymbol{g} = \begin{pmatrix}R^2 & 0 \\ 0 & R^2\sin^2\theta\end{pmatrix}. \tag{2.21}$$

The 'two-volume' or area A of the two-sphere in Figure 2.2 can then be written

$$A = \int_0^{2\pi}\mathrm{d}\phi\int_0^{\pi}\mathrm{d}\theta\,\sqrt{\det\boldsymbol{g}} = \int_0^{2\pi}\mathrm{d}\phi\int_0^{\pi}\mathrm{d}\theta\,R^2\sin\theta = 4\pi R^2, \tag{2.22}$$

as expected.

In Euclidean three-space parallel lines of infinite length never cross, but this could not be proved in Euclidean geometry, so it had to be asserted without proof, the *parallel axiom*. The two-sphere belongs to the class of Riemannian curved spaces which are locally flat: a small portion of the surface can be approximated by its tangential plane. Lines in this plane which are parallel locally do cross when extended far enough, as required for geodesics on the surface of a sphere.

Gaussian Curvature. The deviation of a curved surface from flatness can also be determined from the length of the circumference of a circle. Choose a point 'P' on the surface and draw the locus corresponding to a fixed distance s from that point. If the surface is flat, a plane, the locus is a circle and s is its radius. On a two-sphere of radius R the locus is also a circle, see Figure 2.2, but the distance s is measured along a geodesic. The angle subtended by s at the centre of the sphere is s/R, so the radius of the circle is $r = R\sin(s/R)$. Its circumference is then

$$C = 2\pi R\sin(s/R) = 2\pi s\left(1 - \frac{s^2}{6R^2} + \cdots\right). \tag{2.23}$$

Carl Friedrich Gauss (1777–1855) discovered an invariant characterizing the curvature of two-surfaces, the *Gaussian curvature* K. Although K can be given by a completely general formula independent of the coordinate system (see, for

example, [1]), it is most simply described in an orthogonal system x, y. Let the radius of curvature along the x-axis be $R_x(x)$ and along the y-axis be $R_y(y)$. Then the Gaussian curvature at the point (x_0, y_0) is

$$K = 1/R_x(x_0)R_y(y_0). \tag{2.24}$$

On a two-sphere $R_x = R_y = R$, so $K = R^{-2}$ everywhere. Inserting this into Equation (2.23) we obtain, in the limit of small s,

$$K = \frac{3}{\pi} \lim_{s \to 0} \left(\frac{2\pi s - C}{s^3} \right). \tag{2.25}$$

This expression is true for any two-surface, and it is in fact the only invariant that can be defined.

Whether we live in three or more dimensions, and whether our space is flat or curved, is really a physically testable property of space. Gauss actually proceeded to investigate this by measuring the angles in a triangle formed by three distant mountain peaks. If space were Euclidean the value would be 180°, but if the surface had positive curvature like a two-sphere the angles would add up to more than 180°. Correspondingly, the angles on a saddle surface with negative curvature would add up to less than 180°. This is illustrated in Figures 2.3 and 2.4. The precision in Gauss's time was, however, not good enough to exhibit any disagreement with the Euclidean value.

Comoving Coordinates. If the two-sphere with surface (2.17) and radius R were a balloon expanding with time, $R = R(t)$, points on the surface of the balloon would find their mutual distances scaled by $R(t)$ relative to a time t_0 when the radius was $R_0 = 1$. An observer located at any one point would see all the other points recede radially. This is exactly how we see distant galaxies except that we are not on a two-sphere but, as we shall see, on a spatially curved three-surface with cosmic scale factor $R(t)$.

Suppose this observer wants to make a map of all the points on the expanding surface. It is then no longer convenient to use coordinates dependent on $R(t)$ as in Equations (2.11) and (2.19), because the map would quickly be outdated. Instead it is convenient to factor out the cosmic expansion and replace R by $R(t)\sigma$, where σ is a dimensionless *comoving* coordinate, thus

$$\mathrm{d}l^2 = R^2(t)(\mathrm{d}\sigma^2 + \sigma^2\,\mathrm{d}\theta^2 + \sigma^2 \sin^2\theta\,\mathrm{d}\phi^2). \tag{2.26}$$

Returning to the space we inhabit, we manifestly observe that there are three spatial coordinates, so our space must have at least one dimension more than a two-sphere. It is easy to generalize from the curved two-dimensional manifold (surface) (2.17) embedded in three-space to the curved three-dimensional manifold (hypersurface)

$$x^2 + y^2 + z^2 + w^2 = R^2 \tag{2.27}$$

of a three-sphere (hypersphere) embedded in Euclidean four-space with coordinates x, y, z and a fourth fictitious space coordinate w.

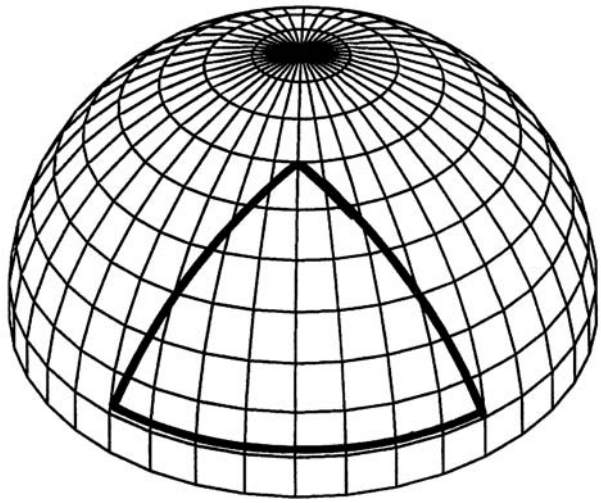


Figure 2.3 The angles in a triangle on a surface with positive curvature add up to more than 180°.

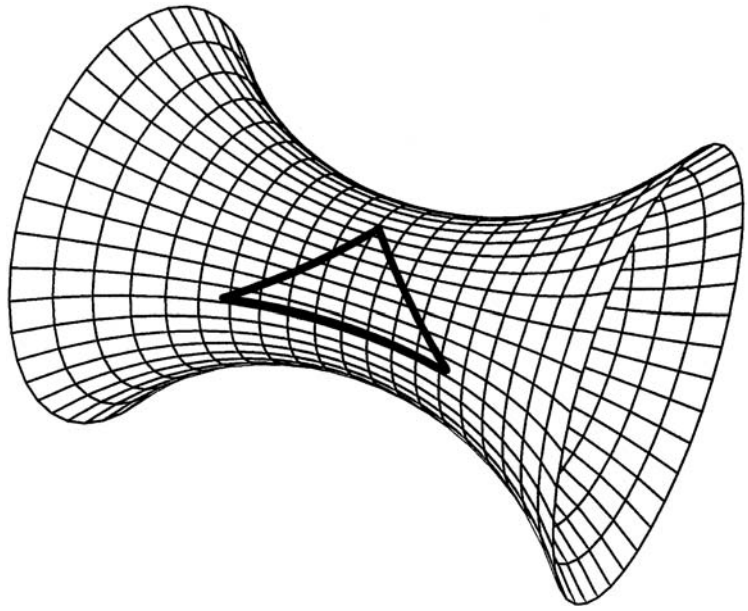


Figure 2.4 The angles in a triangle on a surface with negative curvature add up to less than 180°.

Just as the metric (2.18) could be written without explicit use of z, the metric on the three-sphere (2.27) can be written without use of w,

$$\mathrm{d}l^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 + \frac{(x\,\mathrm{d}x + y\,\mathrm{d}y + z\,\mathrm{d}z)^2}{R^2 - x^2 - y^2 - z^2}, \tag{2.28}$$

or, in the more convenient spherical coordinates used in (2.26),

$$\mathrm{d}l^2 = R^2(t)\left(\frac{R^2\,\mathrm{d}\sigma^2}{R^2-(R\sigma)^2} + \sigma^2\mathrm{d}\theta^2 + \sigma^2\sin^2\theta\,\mathrm{d}\phi^2\right). \tag{2.29}$$

Note that the introduction of the comoving coordinate σ in Equation (2.26) does not affect the parameter R defining the hypersurface in Equation (2.27). No point is preferred on the manifold (2.27), and hence it can describe a spatially homogeneous and isotropic three-dimensional universe in accord with the cosmological principle.

Another example of a curved Riemannian two-space is the surface of a hyperboloid obtained by changing the sign of R^2 in Equation (2.17). The geodesics are hyperbolas, the surface is also unbounded, but in contrast to the spherical surface it is infinite in extent. It can also be generalized to a three-dimensional curved hypersurface, a three-hyperboloid, defined by Equation (2.27) with R^2 replaced by $-R^2$.

The Gaussian curvature of all geodesic three-surfaces in Euclidean four-space is

$$K = k/R^2, \tag{2.30}$$

where the *curvature parameter* k can take the values $+1$, 0, -1, corresponding to the three-sphere, flat three-space, and the three-hyperboloid, respectively. Actually, k can take any positive or negative value, because we can always rescale σ to take account of different values of k.

The Robertson–Walker Metric. Let us now include the time coordinate t and the curvature parameter k in Equation (2.28). We then obtain the complete metric derived by *Howard Robertson* and *Arthur Walker* in 1934:

$$\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 - \mathrm{d}l^2 = c^2\,\mathrm{d}t^2 - R(t)^2\left(\frac{\mathrm{d}\sigma^2}{1-k\sigma^2} + \sigma^2\mathrm{d}\theta^2 + \sigma^2\sin^2\theta\,\mathrm{d}\phi^2\right). \tag{2.31}$$

In the tensor notation of Equation (2.13) the components of the Robertson–Walker metric $\boldsymbol{g}$ are obviously

$$g_{00} = 1, \qquad g_{11} = -\frac{R^2}{1-k\sigma^2}, \qquad g_{22} = -R^2\sigma^2, \qquad g_{33} = -R^2\sigma^2\sin^2\theta. \tag{2.32}$$

Thus if the Universe is homogeneous and isotropic at a given time and has the Robertson–Walker metric (2.32), then *it will always remain homogeneous and isotropic*, because a galaxy at the point (σ, θ, ϕ) will always remain at that point, only the scale of spatial distances $R(t)$ changing with time. The displacements will be $\mathrm{d}\sigma = \mathrm{d}\theta = \mathrm{d}\phi = 0$ and the metric equation will reduce to

$$\mathrm{d}s^2 = c^2\,\mathrm{d}t^2. \tag{2.33}$$

For this reason one calls such an expanding frame a *comoving frame*. An observer at rest in the comoving frame is called a *fundamental observer*. If the Universe appears to be homogeneous to him/her, it must also be isotropic. But another

observer located at the same point and in relative motion with respect to the fundamental observer does not see the Universe as isotropic. Thus the comoving frame is really a preferred frame, and a very convenient one, as we shall see later in conjunction with the cosmic background radiation. Let us note here that a fundamental observer may find that not all astronomical bodies recede radially; a body at motion relative to the comoving coordinates (σ, θ, ϕ) will exhibit peculiar motion in other directions.

Another convenient comoving coordinate is χ, defined by integrating over

$$\mathrm{d}\chi = \frac{\mathrm{d}\sigma}{\sqrt{1 - k\sigma^2}}. \tag{2.34}$$

Inserting this into Equation (2.31), the metric can be written

$$\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 - R^2(t)[\mathrm{d}\chi^2 + S_k^2(\chi)(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)], \tag{2.35}$$

where

$$S_k(\chi) \equiv \sigma$$

and

$$S_1(\chi) = \sin\chi, \qquad S_0(\chi) = \chi, \qquad S_{-1}(\chi) = \sinh\chi. \tag{2.36}$$

We shall use the metrics (2.31) and (2.35) interchangeably since both offer advantages. During the course of gravitational research many other metrics with different mathematical properties have been studied, but none appears to describe the Universe well. Since they are not needed, we shall not discuss them here.

Let us briefly digress to define what is sometimes called *cosmic time*. In an expanding universe the galaxies are all moving away from each other (let us ignore peculiar velocities) with clocks running at different local time, but from our vantage point we would like to have a time value applicable to all of them. If one postulates, with *Hermann Weyl* (1885–1955), that the expansion is so regular that the world lines of the galaxies form a nonintersecting and diverging three-bundle of geodesics, and that one can define spacelike hypersurfaces which are orthogonal to all of them. Then each such hypersurface can be labelled by a constant value of the time coordinate x^0, and using this value one can meaningfully talk about cosmic epochs for the totality of the Universe. This construction in space-time does not imply the choice of a preferred time in conflict with special relativity.

2.3 Relativistic Distance Measures

Let us consider how to measure distances in our comoving frame in which we are at the origin. The *comoving distance* from us to a galaxy at comoving coordinates $(\sigma, 0, 0)$ is not an observable because a distant galaxy can only be observed by the light it emitted at an earlier time $t < t_0$. In a space-time described by the Robertson–Walker metric the light signal propagates along the geodesic $\mathrm{d}s^2 = 0$. Choosing $\mathrm{d}\theta^2 = \mathrm{d}\phi^2 = 0$, it follows from Equation (2.35) that this geodesic is defined by

$$c^2\,\mathrm{d}t^2 - R(t)^2\,\mathrm{d}\chi^2 = 0.$$

It follows that χ can be written

$$\chi = c \int_t^{t_0} \frac{\mathrm{d}t}{R(t)}. \tag{2.37}$$

The time integral in Equation (2.37) is called the *conformal time*.

Proper Distance. Let us now define the *proper distance* d_P at time t_0 (when the scale is R_0) to the galaxy at $(\sigma, 0, 0)$. This is a function of σ and of the intrinsic geometry of space-time and the value of k. Integrating the spatial distance $\mathrm{d}l \equiv |\mathrm{d}\boldsymbol{l}|$ in Equation (2.31) from 0 to d_P we find

$$d_\mathrm{P} = R_0 \int_0^\sigma \frac{\mathrm{d}\sigma}{\sqrt{1-k\sigma^2}} = R_0 \frac{1}{\sqrt{k}} \sin^{-1}(\sqrt{k}\sigma) = R_0\chi. \tag{2.38}$$

For flat space $k = 0$ we find the expected result $d_\mathrm{P} = R_0\sigma$. In a universe with curvature $k = +1$ and scale R the equation (2.38) becomes

$$d_\mathrm{P} = R\chi = R \sin^{-1}\sigma \quad \text{or} \quad \sigma = \sin(d_\mathrm{P}/R).$$

As the distance d_P increases from 0 to $\frac{1}{2}\pi R$, σ also increases from 0 to its maximum value 1. However, when d_P increases from $\frac{1}{2}\pi R$ to πR, σ decreases back to 0. Thus, travelling a distance $d_\mathrm{P} = \pi R$ through the curved three-space brings us to the other end of the Universe. Travelling on from $d_\mathrm{P} = \pi R$ to $d_\mathrm{P} = 2\pi R$ brings us back to the point of departure. In this sense a universe with positive curvature is closed.

Similarly, the area of a three-sphere centred at the origin and going through the galaxy at σ is

$$A = 4\pi R^2\sigma^2 = 4\pi R^2 \sin^2(d_\mathrm{P}/R). \tag{2.39}$$

Clearly, A goes through a maximum when $d_\mathrm{P} = \frac{1}{2}\pi R$, and decreases back to 0 when d_P reaches πR. Note that $A/4$ equals the area enclosed by the circle formed by intersecting a two-sphere of radius R with a horizontal plane, as shown in Figure 2.2. The intersection with an equatorial plane results in the circle enclosing maximal area, $A/4 = \pi R^2$, all other intersections making smaller circles. A plane tangential at either pole has no intersection, thus the corresponding 'circle' has zero area.

The volume of the three-sphere (2.27) can then be written in analogy with Equation (2.22),

$$V = 2\int_0^{2\pi} \mathrm{d}\phi \int_0^{\pi} \mathrm{d}\theta \int_0^1 \mathrm{d}\sigma \sqrt{\det \boldsymbol{g}_\mathrm{RW}}, \tag{2.40}$$

where the determinant of the spatial part of the Robertson–Walker metric matrix $\boldsymbol{g}_\mathrm{RW}$ is now

$$\det \boldsymbol{g}_\mathrm{RW} = R^6 \frac{\sigma^4}{1-\sigma^2} \sin^2\theta. \tag{2.41}$$

The factor 2 in Equation (2.40) comes from the sign ambiguity of w in Equation (2.27). Both signs represent a complete solution. Inserting Equation (2.41) into Equation (2.40) one finds the volume of the three-sphere:

$$V = 2\pi^2 R^3. \tag{2.42}$$

The hyperbolic case is different. Setting $k = -1$, the function in Equation (2.38) is $\mathrm{i}^{-1} \sin^{-1} \mathrm{i}\sigma \equiv \sinh^{-1} \sigma$, thus

$$d_{\mathrm{P}} = R\chi = R \sinh^{-1} \sigma \quad \text{or} \quad \sigma = \sinh(d_{\mathrm{P}}/R). \tag{2.43}$$

Clearly this space is open because σ grows indefinitely with d_{P}. The area of the three-hyperboloid through the galaxy at σ is

$$A = 4\pi R^2 \sigma^2 = 4\pi R^2 \sinh^2(d_{\mathrm{P}}/R). \tag{2.44}$$

Let us differentiate d_{P} in Equation (2.38) with respect to time, noting that σ is a constant since it is a comoving coordinate. We then obtain the Hubble flow v experienced by a galaxy at distance d_{P}:

$$v = \dot{d}_{\mathrm{P}} = \dot{R}(t) \int_0^{\sigma} \frac{\mathrm{d}\sigma}{\sqrt{1 - k\sigma^2}} = \frac{\dot{R}(t)}{R(t)} d_{\mathrm{P}}. \tag{2.45}$$

Thus the Hubble flow is proportional to distance, and Hubble's law emerges in a form more general than Equation (1.19):

$$H(t) = \frac{\dot{R}(t)}{R(t)} = \frac{\dot{a}(t)}{a(t)}. \tag{2.46}$$

Recall that v is the velocity of expansion of the space-time geometry. A galaxy with zero comoving velocity would appear to have a radial recession velocity v because of the expansion.

Particle and Event Horizons. In Equation (1.14) we defined the Hubble radius r_{H} as the distance reached in one Hubble time, τ_{H}, by a light signal propagating along a straight line in flat, static space. Let us define the *particle horizon* σ_{ph} or χ_{ph} (also *object horizon*) as the largest comoving spatial distance from which a light signal could have reached us if it was emitted at time $t = t_{\mathrm{min}} < t_0$. Thus it delimits the size of that part of the Universe that has come into causal contact since time t_{min}. If the past time t is set equal to the last scattering time (the time when the Universe became transparent to light, and thus the earliest time anything was visible, as we will discuss in a later chapter) the particle horizon delimits the visible Universe. From Equation (2.37),

$$\chi_{\mathrm{ph}} = c \int_{t_{\mathrm{min}}}^{t_0} \frac{\mathrm{d}t}{R(t)}, \tag{2.47}$$

and from the notation in Equation (2.36),

$$\sigma_{\mathrm{ph}} = S_k(\chi_{\mathrm{ph}}).$$

A particle horizon exists if t_{min} is in the finite past. Clearly the value of σ_{ph} depends sensitively on the behaviour of the scale of the Universe at that time, $R(t_{\mathrm{ph}})$.

If $k \geqslant 0$, the proper distance (subscript 'P') to the particle horizon (subscript 'ph') at time t is

$$d_{\mathrm{P,ph}} = R(t)\chi_{\mathrm{ph}}. \tag{2.48}$$

Note that d_{P} equals the Hubble radius $r_{\mathrm{H}} = c/H_0$ when $k = 0$ and the scale is a constant, $R(t) = R$. When $k = -1$ the Universe is open, and $d_{\mathrm{P,ph}}$ cannot be interpreted as a measure of its size.

In an analogous way, the comoving distance σ_{eh} to the *event horizon* is defined as the spatially most distant present event from which a world line can ever reach our world line. By 'ever' we mean a finite future time, $t_{\max}$:

$$\chi_{\mathrm{eh}} \equiv c \int_{t_0}^{t_{\max}} \frac{\mathrm{d}t}{R(t)}. \tag{2.49}$$

The particle horizon σ_{ph} at time t_0 lies on our past light cone, but with time our particle horizon will broaden so that the light cone at t_0 will move inside the light cone at $t > t_0$ (see Figure 2.1). The event horizon at this moment can only be specified given the time distance to the ultimate future, $t_{\max}$. Only at $t_{\max}$ will our past light cone encompass the present event horizon. Thus the event horizon is our ultimate particle horizon. Comoving bodies at the particle horizon recede with velocity $c = Hd_{\mathrm{P,ph}}$, but the particle horizon itself recedes even faster. From

$$\mathrm{d}(Hd_{\mathrm{P,ph}})/\mathrm{d}t = \dot{H}d_{\mathrm{P,ph}} + H\dot{d}_{\mathrm{P,ph}} = 0,$$

and making use of the *deceleration parameter* q, defined by

$$q = -\frac{a\ddot{a}}{\dot{a}^2} = -\frac{\ddot{a}}{aH^2}, \tag{2.50}$$

one finds

$$\dot{d}_{\mathrm{P,ph}} = c(q + 1). \tag{2.51}$$

Thus when the particle horizon grows with time, bodies which were at spacelike distances at earlier times enter into the light cone.

The integrands in Equations (2.47) and (2.49) are obviously the same, only the integration limits are different, showing that the two horizons correspond to different conformal times. If $t_{\min} = 0$, the integral in Equation (2.47) may well diverge, in which case there is no particle horizon. Depending on the future behaviour of $R(t)$, an event horizon may or may not exist. If the integral diverges as $t \to \infty$, every event will sooner or later enter the event horizon. The event horizon is then a function of waiting time only, but there exists no event horizon at $t = \infty$. But if $R(t)$ accelerates, so that distant parts of the Universe recede faster than light, then there will be an ultimate event horizon. We shall see later that $R(t)$ indeed appears to accelerate.

Redshift and Proper Distance. In Equation (1.19) in the previous chapter we parametrized the rate of expansion $\dot{a}$ by the Hubble constant H_0. It actually appeared as a dynamical parameter in the lowest-order Taylor expansion of $R(t)$, Equation (1.17). If we allow $H(t)$ to have some mild time dependence, that would correspond to introducing another dynamical parameter along with the next term in the Taylor expansion. Thus adding the second-order term to Equation (1.17), we have for $R(t)$,

$$R(t) \approx R_0 - \dot{R}_0(t_0 - t) + \tfrac{1}{2}\ddot{R}_0(t_0 - t)^2. \tag{2.52}$$

Making use of the definition in Equation (2.46), the second-order expansion for the dimensionless scale factor is

$$a(t) \approx 1 - H_0(t_0 - t) + \tfrac{1}{2}\dot{H}_0(t_0 - t)^2. \tag{2.53}$$

As long as the observational information is limited to R_0 and its first time derivative $\dot{R}_0$, no further terms can be added to these expansions. To account for $\ddot{R}_0$, we shall now make use of the present value of the deceleration parameter (2.50), q_0. Then the lowest-order expression for the cosmological redshift, Equation (1.18), can be replaced by

$$\begin{aligned} z(t) &= (a(t)^{-1} - 1) \\ &= [1 - H_0(t - t_0) - \tfrac{1}{2}q_0H_0^2(t - t_0)^2]^{-1} - 1 \\ &\approx H_0(t - t_0) + (1 + \tfrac{1}{2}q_0)H_0^2(t - t_0)^2. \end{aligned}$$

This expression can further be inverted to express $H_0(t - t_0)$ as a function of the redshift to second order:

$$H_0(t_0 - t) \approx z - (1 + \tfrac{1}{2}q_0)z^2. \tag{2.54}$$

Let us now find the proper distance d_P to an object at redshift z in this approximation. Eliminating χ in Equations (2.37) and (2.38) we have

$$d_P = cR_0 \int_t^{t_0} \frac{\mathrm{d}t}{R(t)} = c \int_t^{t_0} \frac{\mathrm{d}t}{a(t)}.$$

We then insert $a(t)$ from Equation (2.53) to lowest order in $t_0 - t$, obtaining

$$d_P \approx c \int_t^{t_0} [1 + H_0(t_0 - t)]\,\mathrm{d}t = c(t_0 - t)[1 + \tfrac{1}{2}H_0(t_0 - t)]. \tag{2.55}$$

Substituting the expression (2.54) into this yields the sought result:

$$d_P(z) \approx \frac{c}{H_0}(z - \tfrac{1}{2}(1 + q_0)z^2). \tag{2.56}$$

The first term on the right gives Hubble's linear law (1.15), and thus the second term measures deviations from linearity to lowest order. The parameter value $q_0 = -1$ obviously corresponds to no deviation. The linear law has been used to determine H_0 from galaxies within the Local Supercluster (LSC). On the other hand, one also observes deceleration of the expansion in the local universe due to the lumpiness of matter. For instance, the local group clearly feels the overdensity of the Virgo cluster at a distance of about 17 Mpc, falling towards it with a peculiar velocity of about 630 km s^{-1} [4]. It has been argued that the peculiar velocities in the LSC cannot be understood without the pull of the neighbouring Hydra–Centaurus supercluster and perhaps a still larger overdensity in the supergalactic plane, a rich cluster (the A3627) nicknamed 'the Great Attractor'.

It should be clear from this that one needs to go to even greater distances, beyond the influences of local overdensities, to determine a value for q_0. Within the LSC it is safe to conclude that only the linear term in Hubble's law is necessary.

Equation (2.56) is the conventional formula, which is a good approximation for small z. The approximation obviously deteriorates as z increases, so that it attains its maximum at $z = 1/1 + q_0$. In Figure 2.5 we plot the function $d_P(z)$ for small values of z.

Redshift and Luminosity Distance. Consider an astronomical object emitting photons isotropically with power or absolute luminosity L. At the *luminosity distance* d_L from the object we observe only the fraction B_s, its surface brightness, given by the inverse-square distance law

$$B_s = \frac{L}{4\pi d_L^2}. \tag{2.57}$$

Let us now find d_L as a function of z in such a way that the Euclidean inverse-square law (2.57) is preserved. If the Universe does not expand and the object is stationary at proper distance d_P, a telescope with area A will receive a fraction $A/4\pi d_P^2$ of the photons. But in a universe characterized by an expansion $a(t)$, the object is not stationary, so the energy of photons emitted at time t_e is redshifted by the factor $(1 + z) = a^{-1}(t_e)$. Moreover, the arrival rate of the photons suffers time dilation by another factor $(1+z)$, often called the *energy effect*. The *apparent brightness* B_a is then given by

$$B_a = \frac{L}{4\pi d_P^2 (1 + z)^2}. \tag{2.58}$$

Equating $B_a = B_s$ one sees that $d_L = d_P(1 + z)$, and making use of the expression (2.56) one obtains

$$d_L(z) \approx \frac{c}{H_0}(z + \tfrac{1}{2}(1 - q_0)z^2). \tag{2.59}$$

In Figure 2.5 we plot the function $d_L(z)$ for small values of z.

Astronomers usually replace L and B by two empirically defined quantities, *absolute magnitude* M of a luminous object and *apparent magnitude* m. The replacement rule is

$$m - M = -5 + 5 \log d_L, \tag{2.60}$$

where d_L is expressed in parsecs (pc) and the logarithm is to base 10. For example, if one knows the distance d_L to a galaxy hosting a supernova, its absolute magnitude M can be obtained from observations of its apparent magnitude m.

Parallax Distance. Some of the measurements of H_0 in Figure 1.2 depend directly on the calibration of local distance indicators which form the first rung of a ladder of distance measurements. The distances to relatively nearby stars can be measured by the *trigonometrical parallax* up to about 30 pc away (see Table A.1 in the appendix for cosmic distances). This is the difference in angular position of a star as seen from Earth when at opposite points in its circumsolar orbit. The *parallax distance* d_p is defined as

$$d_p = d_P/\sqrt{1 - k\sigma^2}. \tag{2.61}$$

It has been found that most stars in the Galaxy for which we know the luminosity from a kinematic distance determination exhibit a relationship between surface temperature or colour and absolute luminosity, the *Hertzsprung–Russell* relation. These stars are called *main-sequence stars* and they sit on a fairly well-defined curve in the temperature–luminosity plot. Temperature can be determined from colour—note that astronomers define colour as the logarithm of the ratio of the apparent brightnesses in the red and the blue wavelength bands. Cool stars with surface temperature around 3000 K are infrared, thus the part of their spectrum which is in the visible is dominantly red. Hot stars with surface temperature around 12 000 K are ultraviolet, thus the part of their spectrum which is in the visible is dominantly blue. The Sun, with a surface temperature of 5700 K, radiates mainly in the visible, thus its colour is a blended white, slightly yellow. Most main-sequence stars like our Sun are in a prolonged state of steady burning of hydrogen into helium.

Once this empirical temperature–luminosity relation is established, it can be used the other way around to derive distances to farther main-sequence stars: from their colour one obtains the luminosity which subsequently determines d_L. By this method one gets a second rung in a ladder of estimates which covers distances within our Galaxy.

Angular Size Distance. Yet another measure of distance is the *angular size distance* d_A. In Euclidean space an object of size D that is at distance d_A will subtend an angle 2θ such that

$$\theta = \tan(D/d_A) \approx D/d_A,$$

where the approximation is good for small θ. This can serve as the definition of d_A in Euclidean space. In general relativity we can still use this equation to define a distance measure d_A. From the metric equation (2.31) the diameter of a source of light at comoving distance σ is $D = R\sigma\theta$, so

$$d_A = D/\theta = R\sigma = RS_k(d_P/R_0). \tag{2.62}$$

This definition preserves the relation between angular size and distance, a property of Euclidean space. But expansion of the Universe and the changing scale factor R means that as proper distance d_P or redshift z increases, the angular diameter distance initially increases but ultimately decreases. Light rays from the object detected by the observer have been emitted when the proper distance to the object, measured at fixed world time, was small. Because the proper distance between observer and source is increasing faster than the speed of light, emitted light in the direction of the observer is initially moving away from the observer.

The redshift dependence of d_A can be found from Equations (2.56) and (2.36) once k is known. In Figure 2.5 we plot d_A for the choice $k = 0$ when

$$d_A = \frac{R}{R_0} d_P = a d_P = \frac{d_P}{1+z}. \tag{2.63}$$

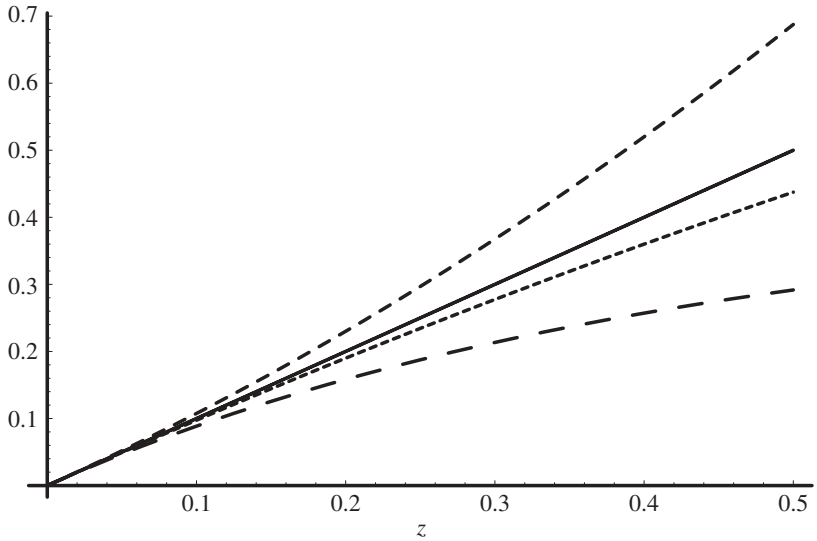


Figure 2.5 Approximate distance measures d to second order in z. The solid curve shows the linear function, Equation (1.15); the short-dashed curve the proper distance d_{P}, Equation (2.56); the medium-dashed curve the luminosity distance d_{L}, Equation (2.59); and the long-dashed curve the angular size distance d_{A}, Equation (2.63). The value of q_0 is -0.5.

The k dependence makes it a useful quantity to determine cosmological parameters. In particular, k is sensitive to certain combinations of well-measured parameters occurring in supernova observations.

Distance Ladder Continued. As the next step on the distance ladder one chooses calibrators which are stars or astronomical systems with specific uniform properties, so called *standard candles*. The *RR Lyrae* stars all have similar absolute luminosities, and they are bright enough to be seen out to about 300 kpc. A very important class of standard candles are the *Cepheid* stars, whose absolute luminosity oscillates with a constant period P in such a way that $\log P \propto 1.3 \log L$. The period P can be observed with good precision, thus one obtains a value for L. Cepheids have been found within our Galaxy where the period–luminosity relation can be calibrated by distances from trigonometric parallax measurements. This permits use of the period–luminosity relation for distances to Cepheids within the *Large Magellanic Cloud* (LMC), our satellite galaxy. At a distance of 55 kpc the LMC is the first important extragalactic landmark.

Globular clusters are gravitationally bound systems of 10^5–10^6 stars forming a spherical population orbiting the centre of our Galaxy. From their composition one concludes that they were created very early in the evolution of the Galaxy. We already made use of their ages to estimate the age of the Universe in Section 1.5. Globular clusters can also be seen in many other galaxies, and they are visible out to 100 Mpc. Within the Galaxy their distance can be measured as described above, and one then turns to study the statistical properties of the clusters: the frequency

of stars of a given luminosity, the mean luminosity, the maximum luminosity, and so on. A well-measured cluster then becomes a standard candle with properties presumably shared by similar clusters at all distances. Similar statistical indicators can be used to calibrate clusters of galaxies; in particular the brightest galaxy in a cluster is a standard candle useful out to 1 Gpc.

The next two important landmarks are the distances to the Virgo cluster, which is the closest moderately rich concentrations of galaxies, and to the Coma cluster, which is one of the closest clusters of high richness. The Virgo distance has been determined to be 17 Mpc by the observations of galaxies containing several Cepheids, by the *Hubble Space Telescope* [5]. The Coma is far enough, about 100 Mpc, that its redshift is almost entirely due to the cosmological expansion.

The existence of different methods of calibration covering similar distances is a great help in achieving higher precision. The expansion can be verified by measuring the surface brightness of standard candles at varying redshifts, the *Tolman test*. If the Universe does indeed expand, the intensity of the photon signal at the detector is further reduced by a factor $(1+z)^2$ due to an optical aberration which makes the surface area of the source appear increased. Such tests have been done and they do confirm the expansion.

The *Tully–Fisher* relation is a very important tool at distances which overlap those calibrated by Cepheids, globular clusters, galaxy clusters and several other methods. This empirical relation expresses correlations between intrinsic properties of whole spiral galaxies. It is observed that their absolute luminosity and their circular rotation velocity v_c are related by

$$L \propto v_c^4. \tag{2.64}$$

The Tully–Fisher relation for spiral galaxies is calibrated by nearby spiral galaxies having Cepheid calibrations, and it can then be applied to spiral galaxies out to 100 Mpc.

For more details on distance measurements the reader is referred to the excellent treatment in the book by Peacock [6].

2.4 General Relativity and the Principle of Covariance

Tensors. In four-dimensional space-time all spatial three-vectors have to acquire a zeroth component just like the line element four-vector ds in Equations (2.6) and (2.13). A vector $\boldsymbol{A}$ with components A^μ in a coordinate system x^μ can be expressed in a transformed coordinate system x'^ν as the vector $\boldsymbol{A}'$ with components

$$A'^\nu = \frac{\partial x'^\nu}{\partial x^\mu} A^\mu, \tag{2.65}$$

where summation over the repeated index μ is implied, just as in Equation (2.13). A vector which transforms in this way is said to be *contravariant*, which is indicated by the *upper index* for the components A^μ.

A vector $\boldsymbol{B}$ with components B_μ in a coordinate system x^μ, which transforms in such a way that

$$B'_\nu = \frac{\partial x^\mu}{\partial x'^\nu} B_\mu, \tag{2.66}$$

is called *covariant*. This is indicated by writing its components with a *lower index*. Examples of covariant vectors are the tangent vector to a curve, the normal to a surface, and the four-gradient of a four-scalar ϕ, $\partial\phi/\partial x^\mu$.

In general, tensors can have several contravariant and covariant indices running over the dimensions of a manifold. In a d-dimensional manifold a tensor with r indices is of *rank* r and has d^r components. In particular, an $r = 1$ tensor is a vector, and $r = 0$ corresponds to a scalar. An example of a tensor is the assembly of the n^2 components $X^\mu Y^\nu$ formed as the products (not the scalar product!) of the n components of the vector X^μ with the n components of the vector Y^ν. We have already met the rank 2 tensors $\eta_{\mu\nu}$ with components given by Equation (2.14), and the metric tensor $g_{\mu\nu}$.

Any contravariant vector $\boldsymbol{A}$ with components A^μ can be converted into a covariant vector by the operation

$$A_\nu = g_{\mu\nu} A^\mu.$$

The contravariant metric tensor $g^{\mu\nu}$ is the matrix inverse of the covariant $g_{\mu\nu}$:

$$g_{\sigma\mu} g^{\mu\nu} = \delta^\nu_\sigma. \tag{2.67}$$

The upper and lower indices of any tensor can be lowered and raised, respectively, by operating with $g_{\mu\nu}$ or $g^{\mu\nu}$ and summing over repeated indices. Thus a covariant vector $\boldsymbol{A}$ with components A_μ can be converted into a contravariant vector by the operation

$$A^\nu = g^{\mu\nu} A_\mu,$$

For a point particle with mass m and total energy

$$E = \gamma m c^2, \tag{2.68}$$

according to Einstein's famous relation, one assigns a momentum four-vector P with components $p^0 = E/c$, $p^1 = p_x$, $p^2 = p_y$, $p^3 = p_z$, so that E and the linear momentum $\boldsymbol{p} = m\boldsymbol{v}$ become two aspects of the same entity, $P = (E/c, \boldsymbol{p})$.

The scalar product P^2 is an invariant related to the mass,

$$P^2 = \eta_{\mu\nu} P^\mu P^\nu = \frac{E^2}{c^2} - p^2 = (\gamma m c)^2, \tag{2.69}$$

where $p^2 \equiv |\gamma \boldsymbol{p}|^2$. For a massless particle like the photon, it follows that the energy equals the three-momentum times c.

Newton's second law in its nonrelativistic form,

$$\boldsymbol{F} = m\boldsymbol{a} = m\dot{\boldsymbol{v}} = \dot{\boldsymbol{p}}, \tag{2.70}$$

is replaced by the relativistic expression

$$F = \frac{\mathrm{d}P}{\mathrm{d}\tau} = \gamma \frac{\mathrm{d}P}{\mathrm{d}t} = \gamma \left(\frac{\mathrm{d}E}{c\,\mathrm{d}t}, \frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} \right). \tag{2.71}$$

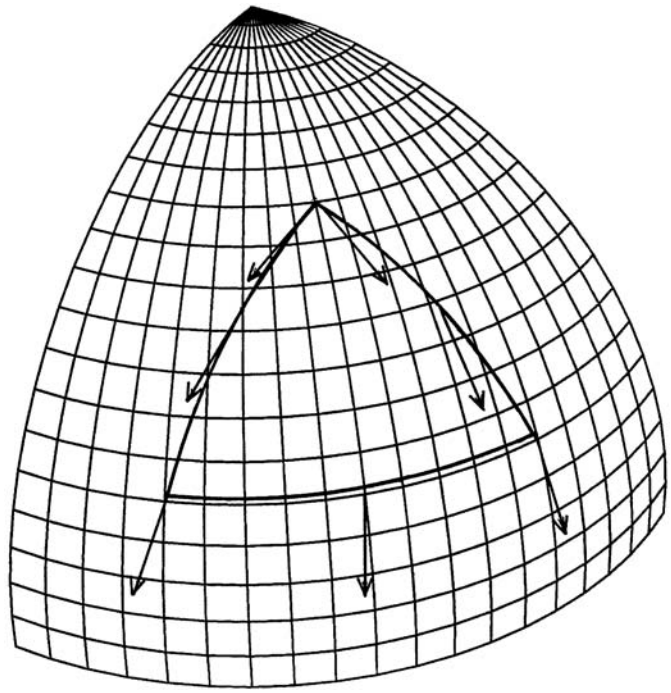


Figure 2.6 Parallel transport of a vector around a closed contour on a curved surface.

General Covariance. Although Newton's second law (2.71) is invariant under special relativity in any inertial frame, it is not invariant in accelerated frames. Since this law explicitly involves acceleration, special relativity has to be generalized somehow, so that observers in accelerated frames can agree on the value of acceleration. Thus the next necessary step is to search for quantities which remain invariant under an arbitrary acceleration and to formulate the laws of physics in terms of these. Such a formulation is called *generally covariant*. In a curved space-time described by the Robertson–Walker metric the approach to general covariance is to find appropriate invariants in terms of tensors which have the desired properties.

Since vectors are rank 1 tensors, vector equations may already be covariant. However, dynamical laws contain many other quantities that are not tensors, in particular space-time derivatives such as $\mathrm{d}/\mathrm{d}\tau$ in Equation (2.71). Space-time derivatives are not invariants because they imply transporting $\mathrm{d}s$ along some curve and that makes them coordinate dependent. Therefore we have to start by redefining derivatives and replacing them with new *covariant derivatives*, which are tensor quantities.

To make the space-time derivative of a vector generally covariant one has to take into account that the direction of a parallel-transported vector changes in terms of the local coordinates along the curve as shown in Figure 2.6. The change is certainly some function of the space-time derivatives of the curved space that is described by the metric tensor.

The covariant derivative operator with respect to the proper time τ is denoted $\mathrm{D}/\mathrm{D}\tau$ (for a detailed derivation, see, for example, references [1] and [2]). Operating with it on the momentum four-vector P^μ results in another four-vector:

$$F^\mu = \frac{\mathrm{D}P^\mu}{\mathrm{D}\tau} \equiv \frac{\mathrm{d}P^\mu}{\mathrm{d}\tau} + \Gamma^\mu_{\sigma\nu} P^\sigma \frac{\mathrm{d}x^\nu}{\mathrm{d}\tau}. \tag{2.72}$$

The second term contains the changes this vector undergoes when it is parallel transported an infinitesimal distance $c\,\mathrm{d}\tau$. The quantities $\Gamma^\mu_{\sigma\nu}$, called *affine connections*, are readily derivable functions of the derivatives of the metric $g_{\mu\nu}$ in curved space-time, but they are not tensors. Their form is

$$\Gamma^\mu_{\sigma\nu} = \tfrac{1}{2} g^{\mu\rho}\left(\frac{\partial g_{\sigma\rho}}{\partial x^\nu} + \frac{\partial g_{\nu\rho}}{\partial x^\sigma} - \frac{\partial g_{\sigma\nu}}{\partial x^\rho}\right). \tag{2.73}$$

With this definition Newton's second law has been made generally covariant.

The path of a test body in free fall follows from Equation (2.72) by requiring that no forces act on the body, $F^\mu = 0$. Making the replacement

$$P^\mu = m\frac{\mathrm{d}x^\mu}{\mathrm{d}\tau},$$

the relativistic equation of motion of the test body, its geodesic, can be written

$$\frac{\mathrm{d}^2 x^\mu}{\mathrm{d}\tau^2} + \Gamma^\mu_{\sigma\nu}\frac{\mathrm{d}x^\sigma}{\mathrm{d}\tau}\frac{\mathrm{d}x^\nu}{\mathrm{d}\tau} = 0. \tag{2.74}$$

In an inertial frame the metric is flat, the metric tensor is a constant everywhere, $g_{\mu\nu}(x) = \eta_{\mu\nu}$, and thus the space-time derivatives of the metric tensor vanish:

$$\frac{\partial g_{\mu\nu}(x)}{\partial x^\rho} = 0. \tag{2.75}$$

It then follows from Equation (2.73) that the affine connections also vanish, and the covariant derivatives equal the simple space-time derivatives.

Going from an inertial frame at x to an accelerated frame at $x + \Delta x$ the expressions for $g_{\mu\nu}(x)$ and its derivatives at x can be obtained as the Taylor expansions

$$g_{\mu\nu}(x + \Delta x) = \eta_{\mu\nu} + \frac{1}{2}\frac{\partial^2 g_{\mu\nu}(x)}{\partial x^\rho \partial x^\sigma}\Delta x^\rho \Delta x^\sigma + \cdots$$

and

$$\frac{\partial g_{\mu\nu}(x + \Delta x)}{\partial x^\rho} = \frac{\partial^2 g_{\mu\nu}(x)}{\partial x^\rho \partial x^\sigma}\Delta x^\sigma + \cdots.$$

The description of a curved space-time thus involves second derivatives of $g_{\mu\nu}$, at least. (Only in a very strongly curved space-time would higher derivatives be needed.)

Recall the definition of the noncovariant Gaussian curvature K in Equation (2.30) defined on a curved two-dimensional surface. In a higher-dimensional space-time, curvature has to be defined in terms of more than just one parameter K. It turns out that curvature is most conveniently defined in terms of the fourth-rank *Riemann tensor*

$$R^\alpha_{\beta\gamma\sigma} = \frac{\partial \Gamma^\alpha_{\beta\sigma}}{\partial x^\gamma} - \frac{\partial \Gamma^\alpha_{\beta\gamma}}{\partial x^\sigma} + \Gamma^\alpha_{\rho\gamma}\Gamma^\rho_{\beta\sigma} - \Gamma^\alpha_{\rho\sigma}\Gamma^\rho_{\beta\gamma}. \tag{2.76}$$

In four-space this tensor has 256 components, but most of them vanish or are not independent because of several symmetries and antisymmetries in the indices. Moreover, an observer at rest in the comoving Robertson–Walker frame will only need to refer to spatial curvature. In a spatial n-manifold, $R^{\alpha}_{\beta\gamma\delta}$ has only $n^2(n^2-1)/12$ nonvanishing components, thus six in the spatial three-space of the Robertson–Walker metric. On the two-sphere there is only one component, which is essentially the Gaussian curvature K.

Another important tool related to curvature is the second rank *Ricci tensor* $R_{\beta\gamma}$, obtained from the Riemann tensor by a summing operation over repeated indices, called *contraction*:

$$R_{\beta\gamma} = R^{\alpha}_{\beta\gamma\alpha} = \delta^{\sigma}_{\alpha} R^{\alpha}_{\beta\gamma\sigma} = g^{\alpha\sigma} R^{\alpha}_{\beta\gamma\sigma}. \tag{2.77}$$

This n^2-component tensor is symmetric in the two indices, so it has only $\frac{1}{2}n(n+1)$ independent components. In four-space the 10 components of the Ricci tensor lead to Einstein's system of 10 gravitational equations as we shall see later. Finally, we may sum over the two indices of the Ricci tensor to obtain the *Ricci scalar* R:

$$R = g^{\beta\gamma} R_{\beta\gamma}, \tag{2.78}$$

which we will need later.

2.5 The Principle of Equivalence

Newton's law of gravitation, Equation (1.28), runs into serious conflict with special relativity in three different ways. Firstly, there is no obvious way of rewriting it in terms of invariants, since it only contains scalars. Secondly, it has no explicit time dependence, so gravitational effects propagate instantaneously to every location in the Universe (in fact, also infinitely far outside the horizon of the Universe!).

Thirdly, the *gravitating mass* m_{G} appearing in Equation (1.28) is totally independent of the *inert mass* m appearing in Newton's second law (2.70), as we already noted, yet for unknown reasons both masses appear to be equal to a precision of 10^{-13} or better (10^{-18} is expected soon). Clearly a theory is needed to establish a formal link between them. Mach thought that the inert mass of a body was somehow linked to the gravitational mass of the whole Universe. Einstein, who was strongly influenced by the ideas of Mach, called this *Mach's principle*. In his early work on general relativity he considered it to be one of the basic, underlying principles, together with the principles of equivalence and covariance, but in his later publications he no longer referred to it.

Facing the above shortcomings of Newtonian mechanics and the limitations of special relativity Einstein set out on a long and tedious search for a better law of gravitation, yet one that would reduce to Newton's law in some limit, of the order of the precision of planetary mechanics.

Weak Principle of Equivalence. Consider the lift in Figure 2.7 moving vertically in a tall tower (it is easy to imagine an lift to be at rest with respect to an outside

observer fixed to the tower, whereas the more 'modern' example of a spacecraft is not at rest when we observe it to be geostationary). A passenger in the lift testing the law of gravitation would find that objects dropped to the floor acquire the usual gravitational acceleration g when the lift stands still, or moves with constant speed. However, when the outside observer notes that the lift is accelerating upwards, tests inside the lift reveal that the objects acquire an acceleration larger than g, and vice versa when the lift is accelerating downwards. In the limit of free fall (unpleasant to the passenger) the objects appear weightless, corresponding to zero acceleration.

Let us now replace the lift with a spacecraft with the engines turned off, located at some neutral point in space where all gravitational pulls cancel or are negligible: a good place is the *Lagrange point*, where the terrestrial and solar gravitational fields cancel. All objects, including the pilot, would appear weightless there.

Now turning on the engines by remote radio control, the spacecraft could be accelerated upwards so that objects on board would acquire an acceleration g towards the floor. The pilot would then rightly conclude that

> *gravitational pull and local acceleration are equivalent*

and indistinguishable if no outside information is available and if $m = m_G$. This conclusion forms the *weak principle of equivalence*, which states that an observer in a gravitational field will not experience free fall as a gravitational effect, but as being at rest in a locally accelerated frame.

A passenger in the lift measuring g could well decide from his local observations that Earth's gravitation actually does not exist, but that the lift is accelerating radially outwards from Earth. This interpretation does not come into conflict with that of another observer on the opposite side of Earth whose frame would accelerate in the opposite direction, because that frame is only local to him/her.

The weak principle of equivalence is already embodied in the *Galilean equivalence principle* in mechanics between motion in a uniform gravitational field and a uniformly accelerated frame of reference. What Einstein did was to generalize this to all of physics, in particular phenomena involving light.

Strong Principle of Equivalence. The more general formulation is the important *strong principle of equivalence* (SPE), which states that

> *to an observer in free fall in a gravitational field the results of all local experiments are completely independent of the magnitude of the field.*

In a suitably small lift or spacecraft, curved space-time can always be approximated by flat Minkowski space-time. In the gravitational field of Earth the gravitational acceleration is directed toward its centre. Thus the two test bodies in Figure 2.8 with a space-like separation do not actually fall along parallels, but along different radii, so that their separation decreases with time. This phenomenon is called the *tidal effect*, or sometimes the tidal force, since the test bodies move as if an attractive exchange force acted upon them. The classic example is the tide caused by the Moon on the oceans. The force experienced by a body of mass m

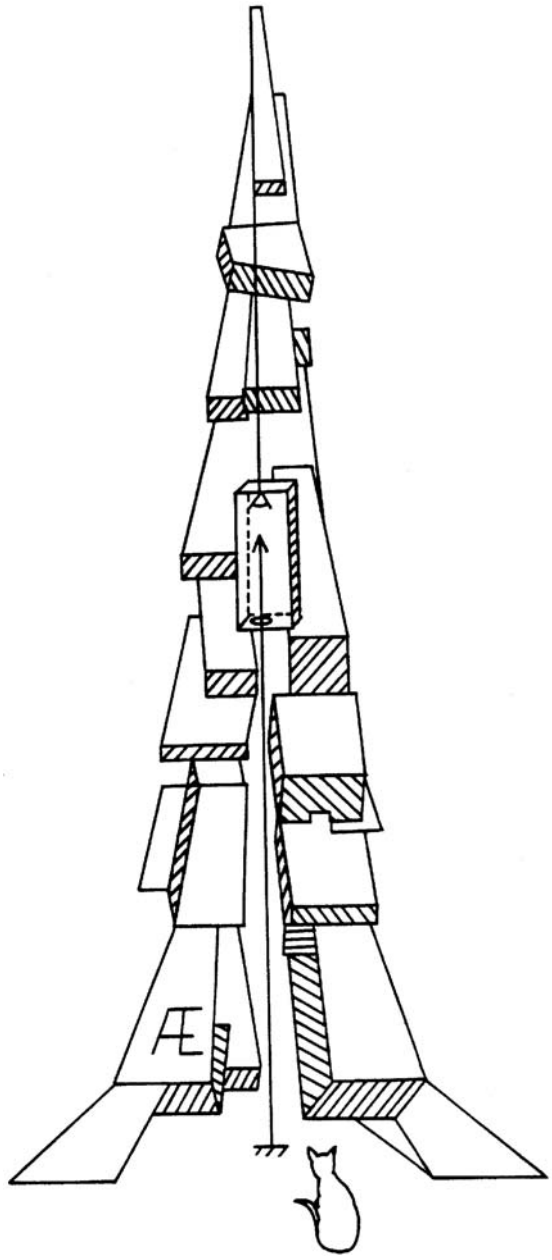


Figure 2.7 The Einstein lift mounted in a non-Euclidean tower. An observer is seen in the foreground.

and diameter d in gravitational interaction with a body of mass M at a distance r is proportional to the differential of the force of attraction (1.28) with respect to r. Neglecting the geometrical shapes of the bodies, the tidal force is

$$F_{\text{tidal}} \approx GMmd/r^3.$$

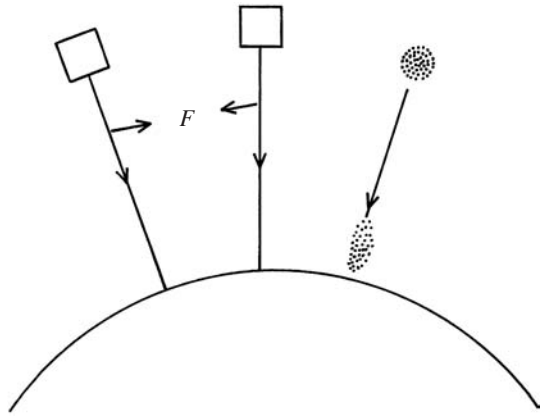


Figure 2.8 Tidal force F acting between two test bodies falling freely towards the surface of a gravitating body. On the right a spherical cluster of small bodies is seen to become ellipsoidal on approaching the body.

Thus parts of m located at smaller distances r feel a stronger force.

An interesting example is offered by a sphere of freely falling particles. Since the strength of the gravitational field increases in the direction of fall, the particles in the front of the sphere will fall faster than those in the rear. At the same time the lateral cross-section of the sphere will shrink due to the tidal effect. As a result, the sphere will be *focused* into an ellipsoid with the same volume. This effect is responsible for the gravitational breakup of very nearby massive stars.

If the tidal effect is too small to be observable, the laboratory can be considered to be local. On a larger scale the gravitational field is clearly quite nonuniform, so if we make use of the principle of equivalence to replace this field everywhere by locally flat frames, we get a patchwork of frames which describe a curved space. Since the inhomogeneity of the field is caused by the inhomogeneous distribution of gravitating matter, Einstein realized that the space we live in had to be curved, and the curvature had to be related to the distribution of matter.

But Einstein had already seen the necessity of introducing a four-dimensional space-time, thus it was not enough to describe space-time in a nonuniform gravitational field by a curved space, time also had to be curved. When moving over a patchwork of local and spatially distinct frames, the local time would also have to be adjusted from frame to frame. In each frame the strong principle of equivalence requires that measurements of time would be independent of the strength of the gravitational field.

Falling Photons. Let us return once more to the passenger in the Einstein lift for a demonstration of the relation between gravitation and the curvature of space-time. Let the lift be in free fall; the passenger would consider that no gravitational field is present. Standing by one wall and shining a pocket lamp horizontally across the lift, she sees that light travels in a straight path, a geodesic in flat space-

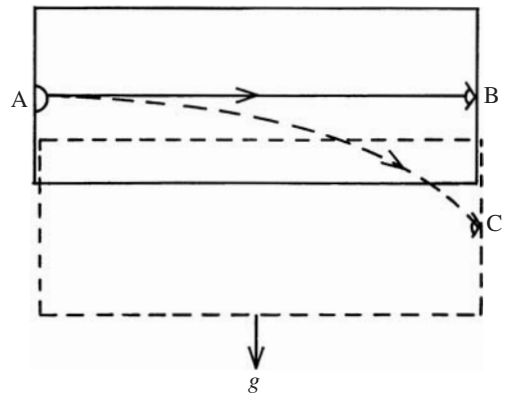


Figure 2.9 A pocket lamp at 'A' in the Einstein lift is shining horizontally on a point 'B'. However, an outside observer who sees that the lift is falling freely with acceleration g concludes that the light ray follows the dashed curve to point 'C'.

time. This is illustrated in Figure 2.9. Thus she concludes that in the absence of a gravitational field space-time is flat.

However, the observer outside the tower sees that the lift has accelerated while the light front travels across the lift, and so with respect to the fixed frame of the tower he notices that the light front follows a curved path, as shown in Figure 2.9. He also sees that the lift is falling in the gravitational field of Earth, and so he would conclude that light feels gravitation as if it had mass. He could also phrase it differently: light follows a geodesic, and since this light path is curved it must imply that space-time is curved in the presence of a gravitational field.

When the passenger shines monochromatic light of frequency ν vertically up, it reaches the roof height d in time d/c. In the same time the outside observer records that the lift has accelerated from, say, $v = 0$ to gd/c, where g is the gravitational acceleration on Earth, so that the colour of the light has experienced a gravitational redshift by the fraction

$$\frac{\Delta\nu}{\nu} \approx \frac{v}{c} = \frac{gd}{c^2} = \frac{GMd}{r^2c^2}. \tag{2.79}$$

Thus the photons have lost energy ΔE by climbing the distance d against Earth's gravitational field,

$$\Delta E = h\Delta\nu = -\frac{gdh\nu}{c^2}, \tag{2.80}$$

where h is the *Planck constant*. (Recall that *Max Planck* (1858–1947) was the inventor of the quantization of energy; this led to the discovery and development of *quantum mechanics*.)

If the pocket lamp had been shining electrons of mass m, they would have lost kinetic energy

$$\Delta E = -gmd \tag{2.81}$$

climbing up the distance d. Combining Equations (2.80) and (2.81) we see that the photons appear to possess mass:

$$m = \frac{h\nu}{c^2}. \tag{2.82}$$

Equation (2.79) clearly shows that light emerging from a star with mass M is redshifted in proportion to M. Thus part of the redshift observed is due to this gravitational effect. From this we can anticipate the existence of stars with so large a mass that their gravitational field effectively prohibits radiation to leave. These are the *black holes* to which we shall return in Section 3.4.

Superluminal Photons. A cornerstone in special relativity is that no material particle can be accelerated beyond c, no physical effect can be propagated faster than c, and no signal can be transmitted faster than c. It is an experimental fact that no particle has been found travelling at superluminal speed, but a name for such particles has been invented, *tachyons*. Special relativity does not forbid tachyons, but if they exist they cannot be retarded to speeds below c. In this sense the speed of light constitutes a two-way barrier: an upper limit for ordinary matter and a lower limit for tachyons.

On quantum scales this may be violated, since the photon may appear to possess a mass caused by its interaction with virtual electron–positron pairs. In sufficiently strong curvature fields, the trajectory of a photon may then be distorted through the interaction of gravity on this mass and on the photon's polarization vector, so that the photon no longer follows its usual geodesic path through curved space-time. The consequence is that SPE may be violated at quantum scales, the photon's lightcone is changed, and it may propagate with superluminal velocity. This effect, called *gravitational birefringence*, can occur because general relativity is not constructed to obey quantum theory. It may still modify our understanding of the origin of the Universe, when the curvature must have been extreme, and perhaps other similar situations like the interior of black holes. For a more detailed discussion of this effect, see Shore [7] and references therein.

2.6 Einstein's Theory of Gravitation

Realizing that the space we live in was not flat, except locally and approximately, Einstein proceeded to combine the principle of equivalence with the requirement of general covariance. The inhomogeneous gravitational field near a massive body being *equivalent* to (a patchwork of flat frames describing) a curved space-time, the laws of nature (such as the law of gravitation) have to be described by *generally covariant* tensor equations. Thus the law of gravitation has to be a covariant relation between mass density and curvature. Because the relativistic field equations cannot be derived, Einstein searched for the simplest form such an equation may take.

The starting point is Newton's law of gravitation, because this has to be true anyway in the limit of very weak fields. From Equation (1.27), the gravitational

force experienced by a unit mass at distance r from a body of mass M and density ρ is a vector in three-space

$$\ddot{\boldsymbol{r}} = \boldsymbol{F} = -\frac{GM\boldsymbol{r}}{r^3},$$

in component form ($i = 1, 2, 3$)

$$\frac{\mathrm{d}^2 x^i}{\mathrm{d}t^2} = F^i = -\frac{GMx^i}{r^3}. \tag{2.83}$$

Let us define a scalar *gravitational potential* ϕ by

$$\frac{\partial \phi}{\partial x^i} = -F^i.$$

This can be written more compactly as

$$\nabla \phi = -\boldsymbol{F}. \tag{2.84}$$

Integrating the flux of the force $\boldsymbol{F}$ through a spherical surface surrounding M and using Stokes's theorem, one can show that the potential ϕ obeys Poisson's equation

$$\nabla^2 \phi = 4\pi G\rho. \tag{2.85}$$

Weak Field Limit. Let us next turn to the relativistic equation of motion (2.74). In the limit of weak and slowly varying fields for which all time derivatives of $g_{\mu\nu}$ vanish and the (spatial) velocity components $\mathrm{d}x^i/\mathrm{d}\tau$ are negligible compared with $\mathrm{d}x^0/\mathrm{d}\tau = c\,\mathrm{d}t/\mathrm{d}\tau$, Equation (2.74) reduces to

$$\frac{\mathrm{d}^2 x^\mu}{\mathrm{d}\tau^2} + c^2 \Gamma^\mu_{00} \left(\frac{\mathrm{d}t}{\mathrm{d}\tau}\right)^2 = 0. \tag{2.86}$$

From Equation (2.73) these components of the affine connection are

$$\Gamma^\mu_{00} = -\tfrac{1}{2} g^{\mu\rho} \frac{\partial g_{00}}{\partial x^\rho},$$

where g_{00} is the time–time component of $g_{\mu\nu}$ and the sum over ρ is implied.

In a weak static field the metric is almost that of flat space-time, so we can approximate $g_{\mu\nu}$ by

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu},$$

where $h_{\mu\nu}$ is a small increment to $\eta_{\mu\nu}$. To lowest order in $h_{\mu\nu}$ we can then write

$$\Gamma^\mu_{00} = -\tfrac{1}{2} \eta^{\mu\rho} \frac{\partial h_{00}}{\partial x^\rho}. \tag{2.87}$$

Inserting this expression into Equation (2.86), the equations of motion become

$$\frac{\mathrm{d}^2 \boldsymbol{x}}{\mathrm{d}\tau^2} = -\frac{1}{2} \left(\frac{\mathrm{d}t}{\mathrm{d}\tau}\right)^2 c^2 \nabla h_{00}, \tag{2.88}$$

$$\frac{\mathrm{d}^2 t}{\mathrm{d}\tau^2} = 0. \tag{2.89}$$

Dividing Equation (2.88) by $(\mathrm{d}t/\mathrm{d}\tau)^2$ we obtain

$$\frac{\mathrm{d}^2\boldsymbol{x}}{\mathrm{d}t^2} = -\tfrac{1}{2}c^2\nabla h_{00}. \tag{2.90}$$

Comparing this with the Newtonian equation of motion (2.83) in the x^i direction we obtain the value of the time–time component of $h_{\mu\nu}$,

$$h_{00} = 2\frac{\phi}{c^2},$$

from which it follows that

$$g_{00} = 1 + 2\frac{\phi}{c^2} = 1 - \frac{2GM}{c^2 r}. \tag{2.91}$$

Stress–Energy Tensor. Let us now turn to the distribution of matter in the Universe. Suppose that matter on some scale can be considered to be continuously distributed as in an *ideal fluid*. The energy density, pressure and shear of a fluid of nonrelativistic matter are compactly described by the *stress–energy tensor* $T_{\mu\nu}$ with the following components.

(i) The time–time component T_{00} is the energy density ρc^2, which includes the mass as well as internal and kinetic energies.

(ii) The diagonal space–space components T_{ii} are the pressure components in the i direction p^i, or the momentum components per unit area.

(iii) The time–space components cT_{0i} are the energy flow components per unit area in the i direction.

(iv) The space–time components cT_{i0} are the momentum densities in the i direction.

(v) The nondiagonal space–space components T_{ij} are the shear components of the pressure p^i in the j direction.

It is important to note that the stress–energy tensor is of rank 2 and is symmetric, thus it has 10 independent components in four-space. However, a comoving observer in the Robertson–Walker space-time following the motion of the fluid sees no time–space or space–time components. Moreover, we can invoke the cosmological principle to neglect the anisotropic nondiagonal space–space components. Thus the stress–energy tensor can be cast into purely diagonal form:

$$T_{\mu\mu} = (p + \rho c^2)\frac{v_\mu v_\mu}{c^2} - p g_{\mu\mu}. \tag{2.92}$$

In particular, the time–time component T_{00} is ρc^2. The conservation of energy and three-momentum, or equivalently the conservation of four-momentum, can be written

$$\frac{\mathrm{D}T_{\mu\nu}}{\mathrm{D}x_\nu} = 0. \tag{2.93}$$

Thus the stress–energy tensor is divergence free.

Taking $T_{\mu\nu}$ to describe relativistic matter, one has to pay attention to its Lorentz transformation properties, which differ from the classical case. Under Lorentz transformations the different components of a tensor do not remain unchanged, but become dependent on each other. Thus the physics embodied by $T_{\mu\nu}$ also differs: the gravitational field does not depend on mass densities alone, but also on pressure. All the components of $T_{\mu\nu}$ are therefore responsible for warping the space-time.

Einstein's Equations. We can now put several things together: replacing ρ in the field equation (2.85) by T_{00}/c^2 and substituting ϕ from Equation (2.91) we obtain a field equation for weak static fields generated by nonrelativistic matter:

$$\nabla^2 g_{00} = \frac{8\pi G}{c^4} T_{00}. \tag{2.94}$$

Let us now assume with Einstein that the right-hand side could describe the source term of a relativistic field equation of gravitation if we made it generally covariant. This suggests replacing T_{00} with $T_{\mu\nu}$. In a matter-dominated universe where the gravitational field is produced by massive stars, and where the pressure between stars is negligible, the only component of importance is then T_{00}.

The left-hand side of Equation (2.94) is not covariant, but it does contain second derivatives of the metric, albeit of only one component. Thus it is already related to curvature. The next step would be to replace $\nabla^2 g_{00}$ by a tensor matching the properties of $T_{\mu\nu}$ on the right-hand side.

(i) It should be of rank two.

(ii) It should be related to the Riemann curvature tensor $R_{\alpha\beta\gamma\sigma}$. We have already found a candidate in the Ricci tensor $R_{\mu\nu}$ in Equation (2.77).

(iii) It should be symmetric in the two indices. This is true for the Ricci tensor.

(iv) It should be divergence-free in the sense of covariant differentiation. This is not true for the Ricci tensor, but a divergence-free combination can be formed with the Ricci scalar R in Equation (2.78):

$$G_{\mu\nu} = R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R. \tag{2.95}$$

$G_{\mu\nu}$ is called the *Einstein tensor*. It contains only terms which are either quadratic in the first derivatives of the metric tensor or linear in the second derivatives.

Thus we arrive at Einstein's covariant formula for the law of gravitation:

$$G_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}. \tag{2.96}$$

The stress–energy tensor $T_{\mu\nu}$ is the sum of the stress–energy tensors for the various components of energy, baryons, radiation, neutrinos, dark matter and possible other forms. Einstein's formula (2.96) expresses that the energy densities, pressures and shears embodied by the stress-energy tensor determine the geometry of space-time, which, in turn, determines the motion of matter.

For weak stationary fields produced by nonrelativistic matter, G_{00} indeed reduces to $\nabla^2 g_{00}$. The Einstein tensor vanishes for flat space-time and in the absence of matter and pressure, as it should. Thus the problems encountered by Newtonian mechanics and discussed at the end of Section 1.6 have been resolved in Einstein's theory. The recession velocities of distant galaxies do not exceed the speed of light, and effects of gravitational potentials are not felt instantly, because the theory is relativistic. The discontinuity of homogeneity and isotropy at the boundary of the Newtonian universe has also disappeared because four-space is unbounded, and because space-time in general relativity is generated by matter and pressure. Thus space-time itself ceases to exist where matter does not exist, so there cannot be any boundary between a homogeneous universe and a void outside space-time.

Problems

1. Starting from the postulates (i) and (ii) in Section 2.1 and the requirement that $\mathrm{d}s^2$ in Equation (2.1) should have the same form in the primed and unprimed coordinates, derive the linear Lorentz transformation (2.2) and the expression (2.3).
2. The radius of the Galaxy is 3×10^{20} m. How fast would a spaceship have to travel to cross it in 300 yr as measured on board? Express your result in terms of $\gamma = 1/\sqrt{1 - v^2/c^2}$ [8].
3. An observer sees a spaceship coming from the west at a speed of $0.6c$ and a spaceship coming from the east at a speed $0.8c$. The western spaceship sends a signal with a frequency of 10^4 Hz in its rest frame. What is the frequency of the signal as perceived by the observer? If the observer sends on the signal immediately upon reception, what is the frequency with which the eastern spaceship receives the signal [8]?
4. If the eastern spaceship in the previous problem were to interpret the signal as one that is Doppler shifted because of the relative velocity between the western and eastern spaceships, what would the eastern spaceship conclude about the relative velocity? Show that the relative velocity must be $(v_1 + v_2)/(1 + v_1 v_2/c^2)$, where v_1 and v_2 are the velocities as seen by an outside observer [8].
5. A source flashes with a frequency of 10^{15} Hz. The signal is reflected by a mirror moving away from the source with speed 10 km s^{-1}. What is the frequency of the reflected radiation as observed at the source [8]?
6. Suppose that the evolution of the Universe is described by a constant decelerating parameter $q = \frac{1}{2}$. We observe two galaxies located in opposite directions, both at proper distance d_P. What is the maximum separation between the galaxies at which they are still causally connected? Express your result as a fraction of distance to d_P. What is the observer's particle horizon?

7. Show that the Hubble distance $r_H = c/H$ recedes with radial velocity

$$\dot{r}_H = c(1 + q). \tag{2.97}$$

8. Is the sphere defined by the Hubble radius r_H inside or outside the particle horizon?
9. Calculate whether the following space-time intervals from the origin are spacelike, timelike or lightlike: (1, 3, 0, 0); (3, 3, 0, 0); (3, -3, 0, 0); (0, 3, 0, 0); (3, 1, 0, 0) [1].
10. The supernova 1987A explosion in the Large Magellanic Cloud 170 000 light years from Earth produced a burst of anti-neutrinos $\bar{\nu}_e$ which were observed in terrestrial detectors. If the anti-neutrinos are massive, their velocity would depend on their mass as well as their energy. What is the proper time interval between the emission, assumed to be instantaneous, and the arrival on Earth? Show that in the limit of vanishing mass the proper time interval is zero. What information can be derived about the anti-neutrino mass from the observation that the energies of the anti-neutrinos ranged from 7 to 11 MeV, and the arrival times showed a dispersion of 7 s?
11. The theoretical resolving power of a telescope is given by $\alpha = 1.22\lambda/D$, where λ is the wavelength of the incoming light and D is the diameter of the mirror. Assuming $D = 5$ m and $\lambda = 8 \times 10^{-7}$ m, determine the largest distance to a star that can be measured by the parallax method. (In reality, atmospheric disturbances set tighter limits.)

Chapter Bibliography

[1] Kenyon, I. R. 1990 *General relativity*. Oxford University Press, Oxford.
[2] Peebles, P. J. E. 1993 *Principles of physical cosmology*. Princeton University Press, Princeton, NJ.
[3] Pyykkö, P. 1988 *Chem. Rev.* **88**, 563.
[4] Lynden-Bell, D. *et al*. 1988 *Astrophys. J.* **326**, 19.
[5] Freedman, W. L. *et al*. 1994 *Nature* **371**, 757.
[6] Peacock, J. A. 1999 *Cosmological physics*. Cambridge University Press, Cambridge.
[7] Shore, G. M. 2002 *Nuclear Phys.* B **633**, 271.
[8] Gasiorowicz, S. 1979 *The structure of matter*. Addison-Wesley, Reading, MA.

3

Gravitational Phenomena

In the previous chapters we have gradually constructed a theory for a possible description of the Universe. However, the theory rests on many assumptions which should be tested before we proceed. For instance, in special relativity spacetime was assumed to be four dimensional and the velocity of light c constant. Moreover, the strong principle of equivalence was assumed to be true in order to arrive at general relativity, and Einstein's law of gravitation, Equation (2.96), is really based on intuition rather than on facts. Consequently, before proceeding to our main task to describe the Universe, we take a look in this chapter at several phenomena predicted by general relativity.

In Section 3.1 we describe the classical tests of general relativity which provided convincing evidence early on that the theory was valid. In Section 3.2 we describe the precision measurements of properties of a binary pulsar which showed convincingly that Einstein was right, and which ruled out most competing theories.

An important gravitational phenomenon is gravitational lensing, encountered already in the early observations of starlight deflected by the passage near the Sun's limb. The lensing of distant galaxies and quasars by interposed galaxy clusters, which is discussed in Section 3.3, has become a tool for studying the internal structure of clusters. Weak lensing is a tool for studying the large-scale distribution of matter in the Universe, and thus for determining some of the cosmological parameters.

If the Einstein equations (2.96) were difficult to derive, it was even more difficult to find solutions to this system of 10 coupled nonlinear differential equations. A particularly simple case, however, is a single star far away from the gravitational influence of other bodies. This is described by the Schwarzschild solution to the Einstein equations in Section 3.4. A particularly fascinating case is a black hole,

Introduction to Cosmology Third Edition by Matts Roos
© 2003 John Wiley & Sons, Ltd ISBN 0 470 84909 6 (cased) ISBN 0 470 84910 X (pbk)

a star of extremely high density. Black holes are certainly the most spectacular prediction of general relativity, and they appear to be ubiquitous in the nuclei of bright galaxies.

The existence of gravitational radiation, already demonstrated in the case of the binary pulsar, is an important prediction of general relativity. However, it remains a great challenge to observe this radiation directly. How to do this will be described in Section 3.5.

3.1 Classical Tests of General Relativity

The classical testing ground of theories of gravitation, Einstein's among them, is celestial mechanics within the Solar System. Ideally one should consider the full many-body problem of the Solar System, a task which one can readily characterize as impossible. Already the relativistic two-body problem presents extreme mathematical difficulties. Therefore, all the classical tests treated only the one-body problem of the massive Sun influencing its surroundings.

The earliest phenomenon requiring general relativity for its explanation was noted in 1859, 20 years before Einstein's birth. The French astronomer *Urban Le Verrier* (1811–1877) found that something was wrong with the planet Mercury's elongated elliptical orbit. As the innermost planet it feels the solar gravitation very strongly, but the orbit is also perturbed by the other planets. The total effect is that the elliptical orbit is nonstationary: it precesses slowly around the Sun. The locus of Mercury's orbit nearest the Sun, the *perihelion*, advances $574''$ (seconds of arc) per century. This is calculable using Newtonian mechanics and Newtonian gravity, but the result is only $531''$, $43''$ too little. Le Verrier, who had already successfully predicted the existence of Neptune from perturbations in the orbit of Uranus, suspected that the discrepancy was caused by a small undetected planet inside Mercury's orbit, which he named Vulcan. That prediction was, however, never confirmed.

With the advent of general relativity the calculations could be remade. This time the discrepant $43''$ were successfully explained by the new theory, which thereby gained credibility. This counts as the first one of three 'classical' tests of general relativity. For details on this test as well as on most of the subsequent tests, see, for example, [1] and [2].

Also, the precessions of Venus and Earth have been put to similar use, and within the Solar System many more consistency tests have been done, based on measurements of distances and other orbital parameters.

The second classical test was the predicted deflection of a ray of light passing near the Sun. We shall come back to that test in Section 3.3 on gravitational lensing.

The third classical test was the gravitational shift of atomic spectra, first observed by *John Evershed* in 1927. The frequency of emitted radiation makes atoms into clocks. In a strong gravitational field these clocks run slower, so the atomic spectra shift towards lower frequencies. This is an effect which we already met in Section 2.5: light emerging from a star with mass M is gravitationally redshifted in proportion to M. Evershed observed the line shifts in a cloud of plasma

ejected by the Sun to an elevation of about 72 000 km above the photosphere and found an effect only slightly larger than that predicted by general relativity. Modern observations of atoms radiating above the photosphere of the Sun have improved on this result, finding agreement with theory at the level of about 2.1×10^{-6}. Similar measurements have been made in the vicinity of more massive stars such as Sirius.

Since then, many experiments have studied the effects of changes in a gravitational potential on the rate of a clock or on the frequency of an electromagnetic signal. Clocks have been put in towers or have travelled in airplanes, rockets and satellites. The so-called 'fourth' test of general relativity, which was conceived by *I. I. Shapiro* in 1964 and carried out successfully in 1971 and later, deserves a special mention. This is based on the prediction that an electromagnetic wave suffers a time delay when traversing an increased gravitational potential.

The fourth test was carried out with the radio telescopes at the Haystack and Arecibo observatories by emitting radar signals towards Mercury, Mars and, notably, Venus, through the gravitational potential of the Sun. The round-trip time delay of the reflected signal was compared with theoretical calculations. Further refinement was achieved later by posing the Viking Lander on the Martian surface and having it participate in the experiment by receiving and retransmitting the radio signal from Earth. This experiment found the ratio of the delay observed to the delay predicted by general relativity to be 1.000 ± 0.002.

Note that the expansion of the Universe and Hubble's linear law (1.15) are not tests of general relativity. Objects observed at wavelengths ranging from radio to gamma rays are close to isotropically distributed over the sky. Either we are close to a centre of spherical symmetry—an anthropocentric view—or the Universe is close to homogeneous. In the latter case, and if the distribution of objects is expanding so as to preserve homogeneity and isotropy (this is local Lorentz invariance), the recession velocities satisfy Hubble's law.

3.2 The Binary Pulsar

The most important tests have been carried out on the radio observations of pulsars that are members of binary pairs, notably the PSR 1913 + 16 discovered in 1974 by *R. A. Hulse* and *J. H. Taylor*, for which they received the Nobel Prize in 1993. Pulsars are rapidly rotating, strongly magnetized neutron stars. If the magnetic dipole axis does not coincide with the axis of rotation (just as is the case with Earth), the star would radiate copious amounts of energy along the magnetic dipole axis. These beams at radio frequencies precess around the axis of rotation like the searchlights of a beacon. As the beam sweeps past our line of sight, it is observable as a pulse with the period of the rotation of the star. Hulse, Taylor and collaborators at Arecibo have demonstrated that pulsars are the most stable clocks we know of in the Universe, the variation is about 10^{-14} on timescales of 6–12 months. The reason for this stability is the intense self-gravity of a neutron star, which makes it almost undeformable until, in a binary pair, the very last few orbits when the pair coalesce into one star.

The pulsar PSR 1913 + 16 is a member of a binary system of two neutron stars which, in addition to their individual spins, rotate around their common centre of mass in a quite eccentric orbit. One of the binary stars is a pulsar, sweeping in our direction with a period of 59 ms, and the binary period of the system is determined to be 7.751 939 337 h. The radial velocity curve as a function of time is known, and from this one can deduce the masses m_1 and m_2 of the binary stars to a precision of 0.0005, as well as the parameters of a Keplerian orbit: the eccentricity and the semi-major axis.

But the system does not behave exactly as expected in Newtonian astronomy, hence the deviations provide several independent confirmations of general relativity. The largest relativistic effect is the apsidal motion of the orbit, which is analogous to the advance of the perihelion of Mercury. A second effect is the counterpart of the relativistic clock correction for an Earth clock. The light travel time of signals from the pulsar through the gravitational potential of its companion provides a further effect.

During 17 years of observations the team has observed a steadily accumulating change of orbital phase of the binary system, which must be due to the loss of orbital rotational energy by the emission of *gravitational radiation*. This rate of change can be calculated since one knows the orbital parameters and the star masses so well. The calculations based on Einstein's general relativity agree to within 1% with the measurements. This is the first observation of gravitational radiation, although it is indirect, since we as yet have no detector with which to receive such waves.

The result is also an important check on competing gravitational theories, several of which have been ruled out. In the near future one can expect even more stringent tests, since there are now several further binary pulsar systems known. Non-pulsating binary systems, which are much more common, are also of interest as sources of high-frequency gravitational radiation to be detected in planned or already existing detectors (Section 3.5).

3.3 Gravitational Lensing

A consequence of the relativistic phenomenon of light rays bending around gravitating masses is that masses can serve as *gravitational lenses* if the distances are right and the gravitational potential is sufficient. Newton discussed the possibility that celestial bodies could deflect light (in 1704), and the astronomer Soldner published a paper (in 1804) in which he obtained the correct Newtonian deflection angle by the Sun, assuming that light was corpuscular in nature. Einstein published the general relativistic calculation of this deflection only in 1936, and it was not until 1979 that the effect was first seen by astronomers.

Recall from Equation (2.82) and Section 2.5 that the Strong Principle of Equivalence (SPE) causes a photon in a gravitational field to move as if it possessed mass. A particle moving with velocity v past a gravitational point potential or a spherically symmetric potential U will experience an acceleration in the transversal direction resulting in a deflection, also predicted by Newtonian dynamics. The

deflection angle α can be calculated from the (negative) potential U by taking the line integral of the transversal gravitational acceleration along the photon's path.

Weak Lensing. In the thin-lens approximation the light ray propagates in a straight line, and the deflection occurs discontinuously at the closest distance. The transversal acceleration in the direction y is then

$$\frac{\mathrm{d}^2 y}{\mathrm{d}t^2} = -\left(1 + \frac{v^2}{c^2}\right)\frac{\mathrm{d}U}{\mathrm{d}y}. \tag{3.1}$$

In Newtonian dynamics the factor in the brackets is just 1, as for velocities $v \ll c$. This is also true if one invokes SPE alone, which accounts only for the distortion of time. However, the full theory of general relativity requires the particle to move along a geodesic in a geometry where space is also distorted by the gravitational field. For photons with velocity c the factor in brackets is then 2, so that the total deflection due to both types of distortion is doubled.

The gravitational distortion can be described as an effective refraction index,

$$n = 1 - \frac{2}{c^2}U > 1, \tag{3.2}$$

so that the speed of light through the gravitational field is reduced to $v = c/n$. Different paths suffer different time delays Δt compared with undistorted paths:

$$\Delta t = \frac{1}{c}\int_{\text{source}}^{\text{observer}} \frac{2}{c^2}\,\mathrm{d}l. \tag{3.3}$$

In Figure 3.1 we show the geometry of a weak lensing event where the potential is weak and the light ray clearly avoids the lensing object. The field equations of GR can then be linearized.

For the special case of a spherically or circularly symmetric gravitating body such as the Sun with mass M inside a radius b, photons passing at distance b of closest approach would be deflected by the angle (Problem 2)

$$\alpha = \frac{4GM}{bc^2}. \tag{3.4}$$

For light just grazing the Sun's limb ($b = 6.96 \times 10^8$ m), the relativistic deflection is $\alpha = 1.750''$, whereas the nonrelativistic deflection would be precisely half of this.

To observe the deflection one needs stars visible near the Sun, so two conditions must be fulfilled. The Sun must be fully eclipsed by the Moon to shut out its intense direct light, and the stars must be very bright to be visible through the solar corona. Soon after the publication of Einstein's theory in 1917 it was realized that such a fortuitous occasion to test the theory would occur on 29 May 1919. The Royal Astronomical Society then sent out two expeditions to the path of the eclipse, one to Sobral in North Brazil and the other one, which included *Arthur S. Eddington* (1882–1944), to the Isle of Principe in the Gulf of Guinea, West Africa.

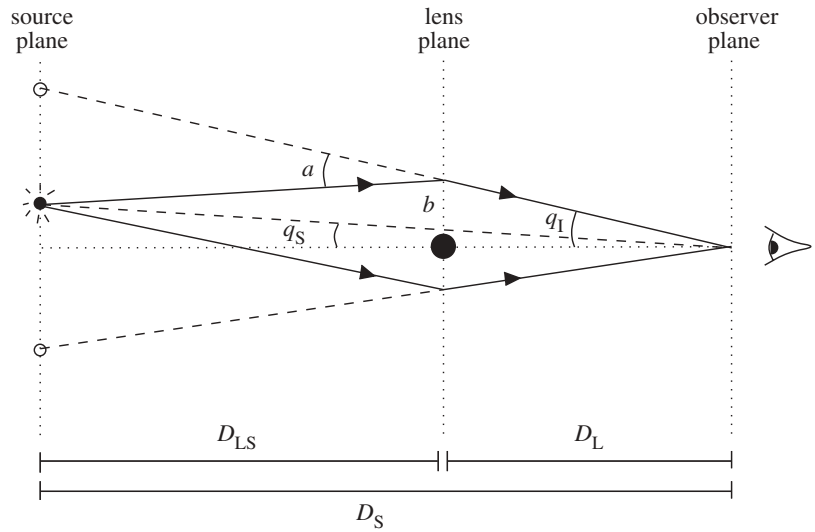


Figure 3.1 The geometry of a gravitational lensing event. For a thin lens, deflection through the small bend angle α may be taken to be instantaneous. The angles θ_I and θ_S specify the observed and intrinsic positions of the source on the sky, respectively. Reprinted from [4] with permission of J. A. Peacock.

Both expeditions successfully observed several stars at various distances from the eclipsed Sun, and the angles of deflection (reduced to the edge of the Sun) were $1.98'' \pm 0.12''$ at Sobral and $1.61'' \pm 0.30''$ at Principe. This confirmed the predicted value of $1.750''$ with reasonable confidence and excluded the Newtonian value of $0.875''$. The measurements have been repeated many times since then during later solar eclipses, with superior results confirming the general relativistic prediction.

The case of starlight passing near the Sun is a special case of a general lensing event, shown in Figure 3.1. For the Sun the distance from lens to observer is small so that the angular size distances are $D_{LS} \approx D_S$, which implies that the actual deflection equals the observed deflection and $\alpha = \theta_I - \theta_S$. In the general case, simple geometry gives the relation between the deflection and the observed displacement as

$$\alpha = \frac{D_S}{D_{LS}}(\theta_I - \theta_S), \tag{3.5}$$

For a lens composed of an ensemble of point masses, the deflection angle is, in this approximation, the vectorial sum of the deflections of the individual point lenses. When the light bending can be taken to be occurring instantaneously (over a short distance relative to D_{LS} and D_L), we have a geometrically thin lens, as assumed in Figure 3.1. Thick lenses are considerably more complicated to analyse.

Strong Lensing. The terms *weak lensing* and *strong lensing* are not defined very precisely. In weak lensing the deflection angles are small and it is relatively easy

to determine the true positions of the lensed objects in the source plane from their displaced positions in the observer plane. Strong lensing implies deflection through larger angles by stronger potentials. The images in the observer plane can then become quite complicated because there may be more than one null geodesic connecting source and observer, so that it is not even always possible to find a unique mapping onto the source plane. Strong lensing is a tool for testing the distribution of mass in the lens rather than purely a tool for testing general relativity.

If a strongly lensing object can be treated as a point mass and is positioned exactly on the straight line joining the observer and a spherical or pointlike lensed object, the lens focuses perfectly and the lensed image is a ring seen around the lens, called an *Einstein ring*. The angular size can be calculated by setting the two expressions for α, Equations (3.4) and (3.5), equal, noting that $\theta_S = 0$ and solving for θ_I:

$$\theta_I = \sqrt{\frac{4GMD_{LS}}{c^2 D_L D_S}}. \tag{3.6}$$

For small M the image is just pointlike. In general, the lenses are galaxy clusters or (more rarely) single galaxies that are not spherical and the geometry is not simple, so that the Einstein ring breaks up into an odd number of sections of arc. Each arc is a complete but distorted picture of the lensed object. Many spectacular lensing pictures can be seen on the HST Internet web page [3].

In general, the solution of the lensing equation and the formation of multiple images can be found by jointly solving Equations (3.4) and (3.5). Equation (3.4) gives the bend angle $\alpha_g(M_b)$ as a function of the gravitational potential for a (symmetric) mass M_b within a sphere of radius b, or the mass seen in projection within a circle of radius b. From Figure 3.1 we can see that $b = \theta_I \times D_{LS}$, so inserting this into Equation (3.4) we have

$$\alpha_g(M_b, \theta_I) = \frac{4GM_b}{c^2 \theta_I D_{LS}}. \tag{3.7}$$

Equation (3.5) is the fundamental lensing equation giving the geometrical relation between the bend angle α_l and the source and image positions:

$$\alpha_l(\theta_S, \theta_I) = \frac{D_S}{D_{LS}}(\theta_I - \theta_S). \tag{3.8}$$

There will be an image at an angle θ_I^* that simultaneously solves both equations:

$$\alpha_g(M_b, \theta_I^*) = \alpha_l(\theta_S, \theta_I^*). \tag{3.9}$$

For the case of a symmetric (or point-mass) lens, θ_I^* will be the two solutions to the quadratic

$$2\theta_I^* = \theta_S + \sqrt{\theta_S^2 + \frac{16GM_b D_{LS}}{c^2 D_L D_S}}. \tag{3.10}$$

This reduces to the radius of the Einstein ring when $\theta_S = 0$. The angle corresponding to the radius of the Einstein ring we denote θ_E.

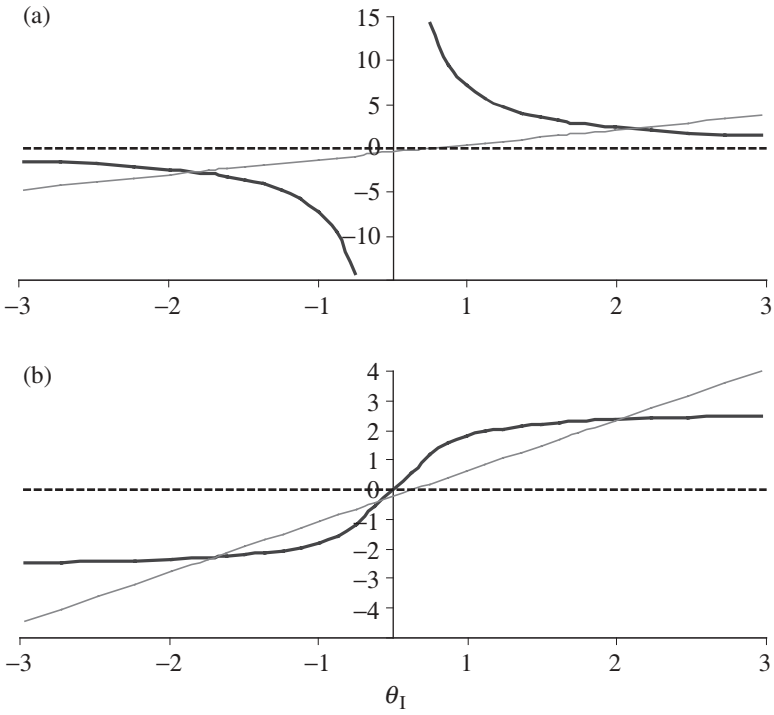


Figure 3.2 Bend angle due to gravitational potential α_g (thick line) and lens geometry α_l (thin line), zero (dashed line). These figures are drawn for a lens of mass $M \approx 7.2\times10^{11}M_\odot$, with $D_S \approx 1.64$ Gpc ($z = 3.4$ in a Friedmann–Robertson–Walker (FRW) cosmology with $\Omega_m = 0.3$ and $\Omega_\lambda = 0.7$), $D_L \approx 1.67$ Gpc ($z = 0.803$), $D_{LS} \approx 0.96$ Gpc and $\alpha_S \approx 0.13''$. (Note that since distances are angular diameter distances, $D_S \neq D_{LS}+D_L$.) In (a) the lens is a point mass and in (b) it is a spherical mass distribution with density given by Equation (3.11) for an ideal galaxy. This roughly corresponds to the parameters associated with the 'Einstein Cross' lensed quasar found by the Hubble telescope, HST 14176 + 5226 [3]. Courtesy of T. S. Coleman.

Graphically, the solutions to the quadratic are given by the intersection of the two curves, $\alpha_g(\theta_I)$ and $\alpha_l(\theta_I)$. The lens equation $\alpha_l(\theta_I)$ (3.8) is a straight line, while the gravitational potential equation $\alpha_g(\theta_I)$ (3.7) depends on the mass distribution. For a point mass, Equation (3.10) describes a pair of hyperbolas and the two curves are as shown in Figure 3.2(a). Clearly there will always be two images for a point-mass lens. When the source displacement is zero ($\theta_S = 0$) the images will be at the positive and negative roots of (3.6)—the Einstein ring. When θ_S is large the positive root will be approximately equal to θ_S, while the negative root will be close to zero (on the line of sight of the lens). This implies that every point-mass lens should have images of every source, no matter what the separation in the sky. Clearly this is not the case. The reason is that the assumption of a point mass and hyperbolic α_g cannot be maintained for small θ_I.

A more realistic assumption for the mass distribution of a galaxy would be that the density is spherically symmetric, with density as a function of distance from

the galactic core, R, given by

$$\rho(R) = \rho_{\text{core}}\left(1 + \frac{R^2}{R_{\text{core}}^2}\right)^{-1}, \tag{3.11}$$

The density is approximately constant (equal to R_{core}) for small radii ($R \ll R_{\text{core}}$) and falls off as R^{-2} for large radii. This roughly matches observed mass-density distributions (including dark matter) as inferred from galaxy rotational-velocity observations (see Section 9.3). The mass will grow like R^3 for $R \ll R_{\text{core}}$ and like R for $R \gg R_{\text{core}}$.

For a spherically symmetric mass with density (3.11), the bend angle due to gravitation, α_{g}, will be as shown in Figure 3.2(b). In this case the straight line given by α_{l} may cross at multiple points, giving multiple images. Figure 3.2(b) shows three solutions, implying three images. In addition, when the mass is not circularly symmetric in the sky, all angles must be treated as two-dimensional vectors on the sky.

For a nonsymmetric mass distribution, the function α_{g} can become quite complicated (see, for example, [5]). Clearly, the problem quickly becomes complex. An example is shown in Figure 3.3, where each light ray from a lensed object propagates as a spherical wavefront. Bending around the lens then brings these wavefronts into positions of interference and self-interaction, causing the observer to see multiple images. The size and shape of the images are therefore changed. From Figure 3.3 one understands how the time delay of pairs of images arises: this is just the time elapsed between different sheets of the same wavefront. In principle, the time delays in Equation (3.3) provide a potential tool for measuring H_0, but not with an interesting precision so far.

Surface Brightness and Microlensing. Since photons are neither emitted nor absorbed in the process of gravitational light deflection, the surface brightness of lensed sources remains unchanged. Changing the size of the cross-section of a light bundle therefore only changes the flux observed from a source and magnifies it at fixed surface-brightness level. For a large fraction of distant quasars the magnification is estimated to be a factor of 10 or more. This enables objects of fainter intrinsic magnitudes to be seen. However, lensing effects are very model dependent, so to learn the true magnification effect one needs very detailed information on the structure of the lens.

If the mass of the lensing object is very small, one will merely observe a magnification of the brightness of the lensed object. This is called *microlensing*, and it has been used to search for nonluminous objects in the halo of our Galaxy. One keeps watch over several million stars in the Large Magellanic Cloud (LMC) and records variations in brightness. Some stars are Cepheids, which have an intrinsic variability, so they have to be discarded. A star which is small enough not to emit visible light and which is moving in the halo is expected to cross the diameter of a star in the LMC in a time span ranging from a few days to a couple of months. The total light amplification for all images from a point-mass lens and point source

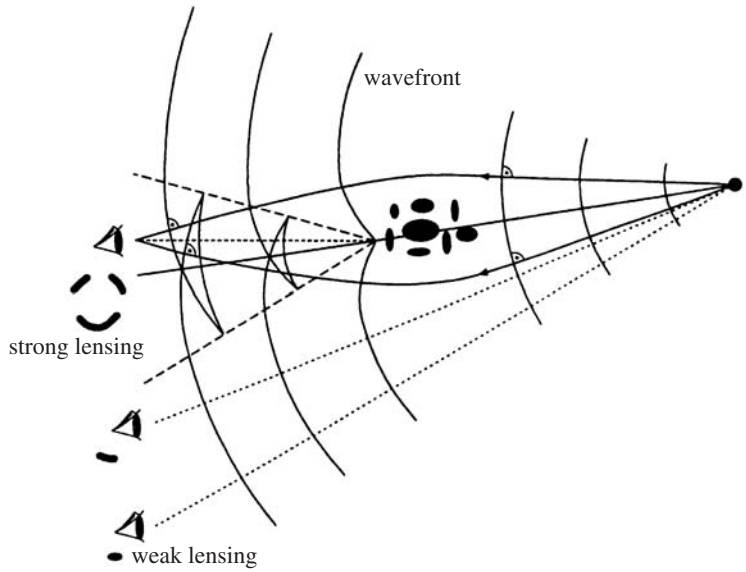


Figure 3.3 Wavefronts and light rays in the presence of a cluster perturbation. Reprinted from [5] with permission of N. Straumann.

is (Problem 3)

$$A = \frac{1 + \frac{1}{2}x}{x\sqrt{1 + \frac{1}{4}x^2}}, \qquad x = \frac{\theta_S}{\theta_E}. \tag{3.12}$$

As the relative positions of the source, lens and observer change, θ_S will change. Simple geometrical arguments give θ_S as a function of the relative velocities, and thus the amplification as a function of time (see [4, pp. 106, 118–120]). During the occultation, the image of the star behind increases in intensity according to this function and subsequently decreases along the time-symmetric curve. A further requirement is that observations in different colours should give the same time curve. Several such microlensing events have been found in the direction of the LMC, and several hundred in the direction of the bulge of our Galaxy. The implications of these discoveries for the search of lumps of nonluminous dark matter will be discussed later.

Cosmic Shear. The large-scale distribution of matter in the Universe is inhomogeneous in every direction, so one can expect that everything we observe is displaced and distorted by weak lensing. Since the tidal gravitational field, and thus the deflection angles, depend neither on the nature of the matter nor on its physical state, light deflection probes the total projected mass distribution. Lensing in infrared light offers the additional advantage of being able to sense distant background galaxies, since their number density is higher than in the optical. The idea of mapping the matter distribution using the *cosmic shear* field was already

proposed (in 1937) by *Fritz Zwicky* (1898–1974), who also proposed looking for lensing by galaxies rather than by stars.

The ray-tracing process mapping a single source into its image can be expressed by the Jacobian matrix between the source-plane coordinates and the observer-plane coordinates:

$$\boldsymbol{J}(\alpha) = \begin{pmatrix} 1-\kappa-\gamma_1 & -\gamma_2 \\ -\gamma_2 & 1-\kappa+\gamma_1 \end{pmatrix}, \tag{3.13}$$

where κ is the *convergence* of the lens and $\gamma = \gamma_1 + \mathrm{i}\gamma_2$ is the shear. The matrix $\boldsymbol{J}(\alpha)$ transforms a circular source into an ellipse with semi-axes stretched by the factor $(1-\kappa \pm |\gamma|)^{-1}$. The convergence affects the isotropic magnification or the projected mass density divided by the critical density, whereas the shear affects the shape of the image. The magnification is given by

$$\mu = (\det \boldsymbol{J})^{-1} = [(1-\kappa)^2 - \gamma^2]^{-1}. \tag{3.14}$$

Clearly, there are locations where μ can become infinite. These points in the source plane are called *caustics* and they lie on the intersections of *critical curves*.

Background galaxies would be ideal tracers of distortions if they were intrinsically circular. Any measured ellipticity would then directly reflect the action of the gravitational tidal field of the interposed lensing matter, and the statistical properties of the cosmic-shear field would reflect the statistical properties of the matter distribution. But many galaxies are actually intrinsically elliptical, and the ellipses are randomly oriented. These intrinsic ellipticities introduce noise into the inference of the tidal field from observed ellipticities.

The sky is covered with a 'wall paper' of faint and distant blue galaxies, about 20 000–40 000 on an area of the size of the full moon. This fine-grained pattern of the sky makes statistical weak-lensing studies possible, because it allows the detection of the coherent distortions imprinted by gravitational lensing on the images of the faint-blue-galaxy population. Several large collaborations carrying out such surveys have reported statistically significant observations of cosmic shear. In the future one can expect large enough surveys to have the sensitivity to identify even invisible (implying X-ray underluminous) galaxy clusters or clumps of matter in the field.

To test general relativity versus alternative theories of gravitation, the best way is to probe the gravitational potential far away from visible matter, and weak galaxy–galaxy lensing is currently the best approach to this end, because it is accurate on scales where all other methods fail, and it is simple if galaxies are treated as point masses. Alternative theories predict an isotropic signal where general relativity predicts an azimuthal variation. The current knowledge is yet preliminary, but favours anisotropy and thus general relativity.

3.4 Black Holes

The Schwarzschild Metric. Suppose that we want to measure time t and radial elevation r in the vicinity of a spherical star of mass M in isolation from all other

gravitational influences. Since the gravitational field varies with elevation, these measurements will surely depend on r. The spherical symmetry guarantees that the measurements will be the same on all sides of the star, and thus they are independent of θ and ϕ. The metric does not then contain $\mathrm{d}\theta$ and $\mathrm{d}\phi$ terms. Let us also consider that we have stable conditions: that the field is static during our observations, so that the measurements do not depend on t.

The metric is then not flat, but the 00 time–time component and the 11 space–space component must be modified by some functions of r. Thus it is of the form

$$\mathrm{d}s^2 = B(r)c^2\,\mathrm{d}t^2 - A(r)\,\mathrm{d}r^2, \tag{3.15}$$

where $B(r)$ and $A(r)$ have to be found by solving the Einstein equations.

Far away from the star the space-time is flat. This gives us the asymptotic conditions

$$\lim_{r\to\infty} A(r) = \lim_{r\to\infty} B(r) = 1. \tag{3.16}$$

From Equation (2.91) the Newtonian limit of g_{00} is known. Here $B(r)$ plays the role of g_{00}; thus we have

$$B(r) = 1 - \frac{2GM}{c^2 r}. \tag{3.17}$$

To obtain $A(r)$ from the Einstein equations is more difficult, and we shall not go to the trouble of deriving it. The exact solution found by *Karl Schwarzschild* (1873–1916) in 1916 preceded any solution found by Einstein himself. The result is simply

$$A(r) = B(r)^{-1}. \tag{3.18}$$

These functions clearly satisfy the asymptotic conditions (3.16).

Let us introduce the concept of *Schwarzschild radius* r_c for a star of mass M, defined by $B(r_\mathrm{c}) = 0$. It follows that

$$r_\mathrm{c} \equiv \frac{2GM}{c^2}. \tag{3.19}$$

The physical meaning of r_c is the following. Consider a body of mass m and radial velocity v attempting to escape from the gravitational field of the star. To succeed, the kinetic energy must overcome the gravitational potential. In the nonrelativistic case the condition for this is

$$\tfrac{1}{2}mv^2 \geqslant GMm/r. \tag{3.20}$$

The larger the ratio M/r of the star, the higher the velocity required to escape is. Ultimately, in the ultra-relativistic case when $v = c$, only light can escape. At that point a nonrelativistic treatment is no longer justified. Nevertheless, it just so happens that the equality in (3.20) fixes the radius of the star correctly to be precisely r_c, as defined above. Because nothing can escape the interior of r_c, not even light, *John A. Wheeler* coined the term black hole for it in 1967. Note that the escape velocity of objects on Earth is 11 km s^{-1}, on the Sun it is 2.2×10^6 km h^{-1}, but on a black hole it is c.

This is the simplest kind of a *Schwarzschild black hole*, and r_c defines its event horizon. Inserting r_c into the functions A and B, the *Schwarzschild metric* becomes

$$d\tau^2 = \left(1 - \frac{r_c}{r}\right) dt^2 - \left(1 - \frac{r_c}{r}\right)^{-1} \frac{dr^2}{c^2}. \tag{3.21}$$

Falling Into a Black Hole. The Schwarzschild metric has very fascinating consequences. Consider a spacecraft approaching a black hole with apparent velocity $v = dr/dt$ in the fixed frame of an outside observer. Light signals from the spacecraft travel on the light cone, $d\tau = 0$, so that

$$\frac{dr}{dt} = c\left(1 - \frac{r_c}{r}\right). \tag{3.22}$$

Thus the spacecraft appears to slow down with decreasing r, finally coming to a full stop as it reaches $r = r_c$.

No information can ever reach the outside observer beyond the event horizon. The reason for this is the mathematical singularity of dt in the expression

$$c\,dt = \frac{dr}{1 - r_c/r}. \tag{3.23}$$

The time intervals dt between successive crests in the wave of the emitted light become longer, reaching infinite wavelength at the singularity. Thus the frequency ν of the emitted photons goes to zero, and the energy $E = h\nu$ of the signal vanishes. One cannot receive signals from beyond the event horizon because photons cannot have negative energy. Thus the outside observer sees the spacecraft slowing down and the signals redshifting until they cease completely.

The pilot in the spacecraft uses local coordinates, so he sees the passage into the black hole entirely differently. If he started out at distance r_0 with velocity $dr/dt = 0$ at time t_0, he will have reached position r at proper time τ, which we can find by integrating $d\tau$ in Equation (3.21) from 0 to τ:

$$\int_0^\tau \sqrt{d\tau^2} = \tau = \int_{r_0}^{r} \left[\frac{1 - r_c/r}{(dr/dt)^2} - \frac{1}{c^2(1 - r_c/r)}\right]^{1/2} dr. \tag{3.24}$$

The result depends on $dr(t)/dt$, which can only be obtained from the equation of motion. The pilot considers that he can use Newtonian mechanics, so he may take

$$\frac{dr}{dt} = c\sqrt{\frac{r_c}{r}}.$$

The result is then (Problem 7)

$$\tau \propto (r_0 - r)^{3/2}. \tag{3.25}$$

However, many other expressions for $dr(t)/dt$ also make the integral in Equation (3.24) converge.

Thus the singularity at r_c does not exist to the pilot, his comoving clock shows finite time when he reaches the event horizon. Once across r_c the spacecraft

reaches the centre of the black hole rapidly. For a hole of mass $10M_\odot$ this final passage lasts about 10^{-4} s. The fact that the singularity at r_c does not exist in the local frame of the spaceship indicates that the horizon at r_c is a mathematical singularity and not a physical singularity. The singularity at the horizon arises because we are using, in a region of extreme curvature, coordinates most appropriate for flat or mildly curved space-time. Alternate coordinates, more appropriate for the region of a black hole and in which the horizon does not appear as a singularity, were invented by Eddington (1924) and rediscovered by Finkelstein (1958) (cf. [1]).

Although this spacecraft voyage is pure science fiction, we may be able to observe the collapse of a supernova into a black hole. Just as for the spacecraft, the collapse towards the Schwarzschild radius will appear to take a very long time. Towards the end of it, the ever-redshifting light will fade and finally disappear completely.

Note from the metric equation (3.21) that inside r_c the time term becomes negative and the space term positive, thus space becomes timelike and time spacelike. The implications of this are best understood if one considers the shape of the light cone of the spacecraft during its voyage (see Figure 3.4). Outside the event horizon the future light cone contains the outside observer, who receives signals from the spacecraft. Nearer r_c the light cone narrows and the slope dr/dt steepens because of the approaching singularity in expression on the the right-hand side of Equation (3.22). The portion of the future space-time which can receive signals therefore diminishes.

Since the time and space axes have exchanged positions inside the horizon, the future light cone is turned inwards and no part of the outside space-time is included in the future light cone. The slope of the light cone is vertical at the horizon. Thus it defines, at the same time, a cone of zero opening angle around the original time axis, and a cone of 180° around the final time axis, encompassing the full space-time of the black hole. As the spacecraft approaches the centre, dt/dr decreases, defining a narrowing opening angle which always contains the centre. When the centre is reached, the spacecraft no longer has a future.

Black Hole Properties. At the centre of the hole, the metric (3.21) is singular. This represents a physical singularity. One cannot define field equations there, so general relativity breaks down, unable to predict what will happen. Some people have speculated that matter or radiation falling in might 'tunnel' through a 'wormhole' out into another universe. Needless to say, all such ideas are purely theoretical speculations with no hope of experimental verification.

A black hole is a region from which nothing can escape, not even light. Black holes are very simple objects as seen from outside their event horizon, they have only the three properties: mass, electric charge and angular momentum. Their size depends only on their mass so that all holes with the same mass are identical and exactly spherical, unless they rotate. All other properties possessed by stars, such as shape, solid surface, electric dipole moment, magnetic moments, as well as any detailed outward structure, are absent. This has led to John Wheeler's famous statement 'black holes have no hair'.

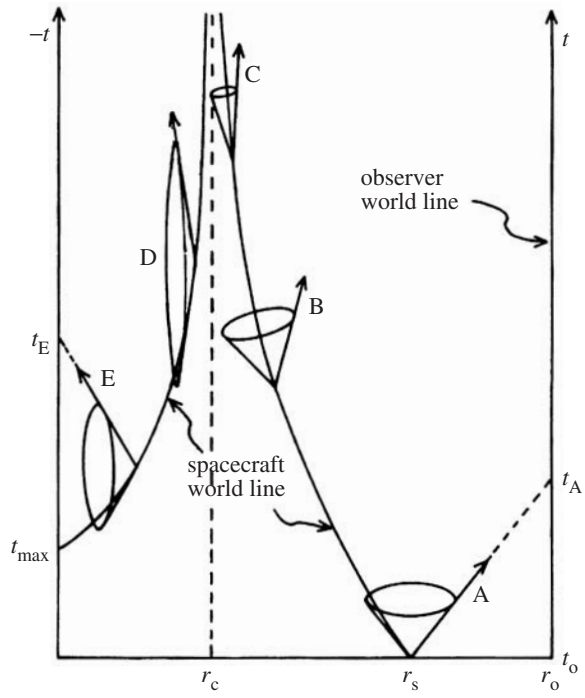


Figure 3.4 The world line of a spacecraft falling into a Schwarzschild black hole. A, the journey starts at time t_0 when the spacecraft is at a radius r_s, far outside the Schwarzschild radius r_c, and the observer is at r_0. A light signal from the spacecraft reaches the observer at time $t_A > t_0$ (read time on the right-hand vertical scale!). B, nearer the black hole the future light cone of the spacecraft tilts inward. A light signal along the arrow will still reach the observer at a time $t_B \gg t_A$. C, near the Schwarzschild radius the light cone narrows considerably, and a light signal along the arrow reaches the observer only in a very distant future. D, inside r_c the time and space directions are interchanged, time running from up to down on the left-hand vertical scale. All light signals will reach the centre of the hole at $r = 0$, and none will reach the observer. The arrow points in the backward direction, so a light signal will reach the centre after the spacecraft. E, the arrow points in the forward direction of the hole, so that a light signal will reach the centre at time t_E, which is earlier than t_{max}, when the spacecraft ends its journey.

Black holes possessing either charge or angular momentum are called *Reissner–Nordström black holes* and *Kerr black holes*, respectively, and they are described by different metrics. It is natural to consider that matter attracted by a hole has angular momentum. Matter can circulate a hole in stable orbits with radii exceeding $3r_c$, but if it comes any closer it starts to spiral in towards the horizon, and is soon lost into the hole with no possibility to escape. Since angular momentum is conserved, the infalling matter must speed up the rotation of the hole. However, centrifugal forces set a limit on the angular momentum J that a rotating black hole can possess:

$$J \leqslant \frac{GM^2}{c}. \tag{3.26}$$

This does not imply that the hole is ripped into pieces with one increment of rotating matter, rather, that it could never have formed in the first place. Remember that angular momentum is energy, and energy is curvature, so incremental energy is modifying the space-time geometry of the black hole, leading to a smaller event horizon. Thus the angular momentum can never overcompensate the gravitational binding energy. If it could, there would be no event horizon and we would have the case of a visible singularity, also called a *naked singularity*. Since nobody has conceived of what such an object would look like, *Stephen Hawking* and *Roger Penrose* have conjectured that space-time singularities should always be shielded from inspection by an event horizon. This is called the principle of *cosmic censorship*—in Penrose's words 'Nature abhors a naked singularity'. The reader might find further enjoyment reading the book by Hawking and Penrose on this subject [6].

J. Bekenstein noted in 1973 [7] that there are certain similarities between the size of the event horizon of a black hole and entropy. When a star has collapsed to the size of its Schwartzschild radius, its event horizon will never change (to an outside observer) although the collapse continues (see Figure 3.4). Thus entropy s could be defined as the surface area A of the event horizon times some proportionality factor,

$$s = \frac{Akc^3}{4G\hbar}, \tag{3.27}$$

the *Bekenstein–Hawking formula*. For a spherically symmetric black hole of mass M the surface area is given by

$$A = 8\pi M^2 G^2/c^2. \tag{3.28}$$

A can increase only if the black hole devours more mass from the outside, but A can never decrease because no mass will leave the horizon. Inserting this into Equation (3.27), entropy comes out proportional to M^2:

$$s = M^2 2\pi kGc/\hbar. \tag{3.29}$$

Thus two black holes coalesced into one possess more entropy than they both had individually. This is illustrated in Figure 3.5.

Hawking Radiation. Stephen Hawking has shown [8, 9] that although no light can escape from black holes, they can nevertheless radiate if one takes quantum mechanics into account. It is a property of the *vacuum* that particle–antiparticle pairs such as $e^- e^+$ are continuously created out of nothing, to disappear in the next moment by *annihilation*, which is the inverse process. Since energy cannot be created or destroyed, one of the particles must have positive energy and the other one an equal amount of negative energy. They form a *virtual pair*, neither one is real in the sense that it could escape to infinity or be observed by us.

In a strong electromagnetic field the electron e^- and the positron e^+ may become separated by a Compton wavelength λ of the order of the Schwarzschild radius r_c. Hawking has shown that there is a small but finite probability for one of

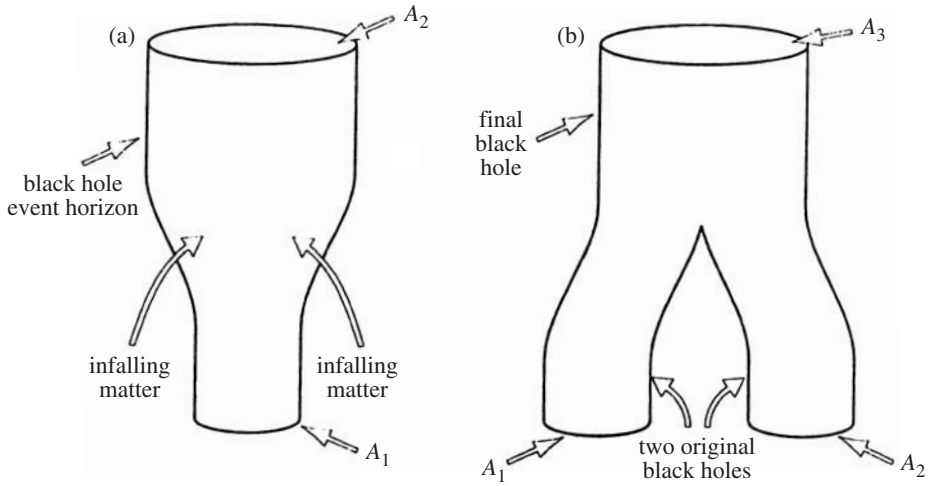


Figure 3.5 When we throw matter into a black hole, or allow two black holes to merge, the total area of the event horizons will never decrease. (a) $A_2 \geqslant A_1$, (b) $A_3 \geqslant A_1 + A_2$. From S. Hawking and R. Penrose [6], copyright 1996 by Princeton University Press. Reprinted by permission of Princeton University Press.

them to 'tunnel' through the barrier of the quantum vacuum and escape the black hole horizon as a real particle with positive energy, leaving the negative-energy particle inside the horizon of the hole. Since energy must be conserved, the hole loses mass in this process, a phenomenon called *Hawking radiation*.

The timescale of complete evaporation is

$$t \approx 10\ \text{Gyr}\left(\frac{M}{10^{12}\ \text{kg}}\right)^3. \tag{3.30}$$

Thus small black holes evaporate fast, whereas heavy ones may have lifetimes exceeding the age of the Universe.

The analogy with entropy can be used even further. A system in thermal equilibrium is characterized by a unique temperature T throughout. When Hawking applied quantum theory to black holes, he found that the radiation emitted from particle–antiparticle creation at the event horizon is exactly thermal. The rate of particle emission is as if the hole were a black body with a unique temperature proportional to the gravitational field on the horizon, the *Hawking temperature*:

$$T_{\text{H}} = \frac{1}{8\pi GM} = 6.15 \times 10^{-8}\frac{M_\odot}{M}\ \text{K}. \tag{3.31}$$

Black Hole Creation. Black holes may have been created in the Big Bang, and they are probably created naturally in the ageing of stars. As explained in Section 1.4, the gravitational collapse of stars burning light elements to heavier elements by nuclear fusion is balanced by the gas pressure. At the end of this cycle they become red giants with iron cores. If the mass of the core does not exceed

the Chandrasekhar limit of $1.4M_\odot$, it is stabilized against gravitational collapse by the electron degeneracy pressure.

But there is a limit to how densely electrons can be packed; if the mass exceeds the Chandrasekhar limit, even the electron degeneracy pressure cannot withstand the huge force of gravity, and the star may collapse and explode as a supernova. At this stage the core consists of neutrons, protons, some electrons and a few neutrinos, forming a *proto-neutron star* that is stabilized against gravity by the degeneracy pressure of the nucleons. Although the theoretical understanding of the collapse and explosion is still poor, it is believed that the proto-neutron star boils violently during its first second, is then flattened by rotation, and finally becomes a spherical neutron star with nuclear density and a solid surface, with a mass in the range 1.4–$1.8M_\odot$ and a radius of $r \approx 3r_\mathrm{c}$.

However, if the mass exceeds the *Landau–Oppenheimer–Volkov limit* of $2.3M_\odot$ (sometimes the limit $1.8M_\odot$ is quoted), there is no stability and no bounce, and the neutron star collapses further to become a black hole.

The fate of a collapsing spherical star can be illustrated schematically by a light cone diagram (see Figure 3.6). Since we cannot draw four-dimensional pictures, we consider only the evolution in time of an event horizon corresponding to the circular equatorial intersection of the star. With increasing time—vertically upwards in the figure—the equatorial intersection shrinks and light rays toward the future steepen. When the star has shrunk to the size of the Schwartzschild radius, the equatorial section has become a trapped surface, the future of all light rays from it are directed towards the world line of the centre of the star. For an outside observer the event horizon then remains eternally of the same size, but an insider would find that the trapped surface of the star is doomed to continue its collapse toward the singularity at zero radius. The singularity has no future light cone and cannot thus ever be observed.

The collapse of an isolated heavy star is, however, not the only route to the formation of a black hole, and probably not even the most likely one. Binary stars are quite common, and a binary system consisting of a neutron star and a red giant is a very likely route to a black hole. The neutron star spirals in, accretes matter from the red giant at a very high rate, about $1M_\odot$ per year, photons are trapped in the flow until the temperature rises above 1 MeV, when neutrinos can carry off the thermal energy, and the neutron star collapses into a black hole.

Black Hole Candidates. Binary systems consisting of a black hole and either a main-sequence star, a neutron star or another hole are likely to give rise to a heavier hole. For example, Cyg X-1 is a black hole—a main-sequence star binary with a hole mass of more than about $10M_\odot$, probably the most massive black hole in a binary observed in the Galaxy. The enormous gravitational pull of the hole tears material from its companion star. This material then orbits the hole in a Saturnus-like accretion disc before disappearing into the hole. Gravity and friction heat the material in the accretion disc until it emits X-rays. Finally, the main-sequence star explodes and becomes a neutron star, which will ultimately merge into the hole to form a heavier black hole.

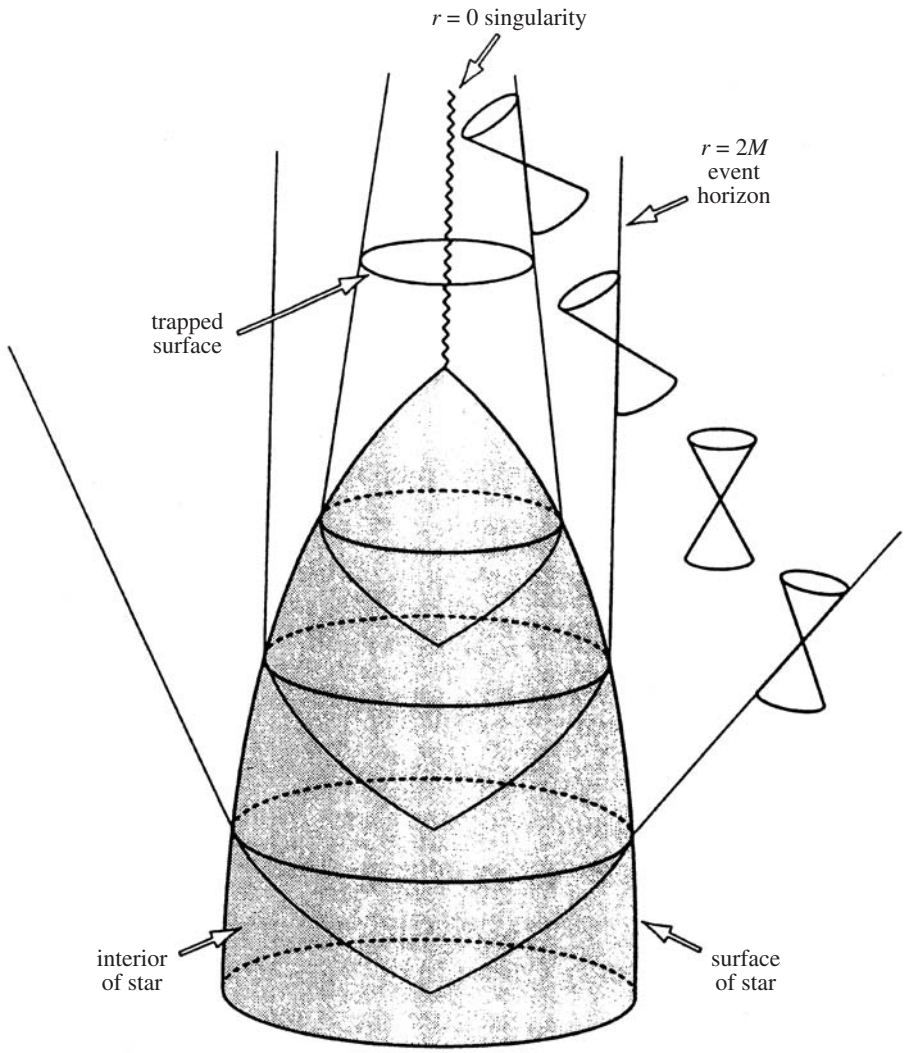


Figure 3.6 A space-time picture of the collapse of a star to form a black hole, showing the event horizon and a closed trapped surface. From S. Hawking and R. Penrose [6], copyright 1996 by Princeton University Press. Reprinted by permission of Princeton University Press.

Although black holes have long been suspected to be the enormously powerful engines in quasars residing in active galactic nuclei (AGN), observational evidence suggests a ubiquity of holes in the nuclei of all bright galaxies, regardless of their activity. Among some 50 hole candidates, the best case is under study in the centre of our Galaxy near the radio source Sgr A$*$. The proof is assembled by measuring the velocity vectors of many stars within 1 pc of Sgr A$*$ and tracing their curved stellar orbits, thereby inferring their acceleration. All the acceleration vectors intersect at Sgr A$*$, and the velocity vectors do not decrease with decreasing

distance to Sgr A∗, indicating that the stars move in a very strong gravitational field of a practically pointlike source of mass $(3.7 \pm 1.5) \times 10^6 M_\odot$. In particular, 10 yr of astrometric imaging has permitted the tracing of two-thirds of a bound highly elliptical Keplerian orbit of the star currently closest to Sgr A∗ [10]. These data no longer allow for a central dense cluster of dark stellar objects or a ball of massive degenerate fermions.

From measurements of the velocities of 64 individual stars in the central region of the dense globular cluster M15 the Hubble Space Telescope inferred in 2002 [11] that M15 must have a central concentration of nonluminous material. If this is due to a single black hole, then its mass is $M = (3.9 \pm 2.2) \times 10^3 M_\odot$, and is thus of intermediate size.

Many other candidates have been spotted by the X-ray signal from the accretion of surrounding matter. A plausible black hole has been found in the spiral galaxy NGC4258 in the constellation Canes Venatici at a distance of 6.4 Mpc [12, 13]. A disk of water vapour (observed by its maser emission) and other molecular material, of mass up to $4 \times 10^6 M_\odot$, is rotating in a Keplerian orbit near the galaxy's nucleus at velocities near 1000 km s^{-1}. Such high velocities for a rotating molecular disk would require the gravitational pull of a black hole of mass $3.9 \times 10^7 M_\odot$. This galaxy also shows other features expected from black holes acting as central engines in active galaxies, such as jets of gas that are twisted into the shape of a helix emerging from the nucleus at speeds of 600 km s^{-1}. It can safely be ruled out that the gravitational field is generated by a cluster of small dark objects such as stellar black holes, neutron stars, etc., because such clusters are expected to have a lifetime of less than 10^8 yr, too short with respect to the Hubble time.

3.5 Gravitational Waves

Einstein noted in 1916 that his general relativity predicted the existence of gravitational radiation, but its possible observation is still in the future. As we explained in Section 3.2, the slowdown of the binary pulsar PSR 1913 + 16 is indirect evidence that this system loses its energy by radiating gravitational waves.

When gravitational waves travel through space-time they produce ripples of curvature, an oscillatory stretching and squeezing of space-time analogous to the tidal effect of the Moon on Earth. Any matter they pass through will feel this effect. Thus a detector for gravitational waves is similar to a detector for the Moon's tidal effect, but the waves act on an exceedingly weaker scale.

Gravitational radiation travels with the speed of light and traverses matter unhindered and unaltered. It may be that the carriers are particles, *gravitons*, with spin $J = 2$, but it is hard to understand how that could be verified. Perhaps, if a theory were found combining gravitation and quantum mechanics, the particle nature of gravitational radiation would be more meaningful.

Tensor Field. In contrast to the electromagnetic field, which is a vector field, the gravitational field is a tensor field. The gravitational analogue of electromag-

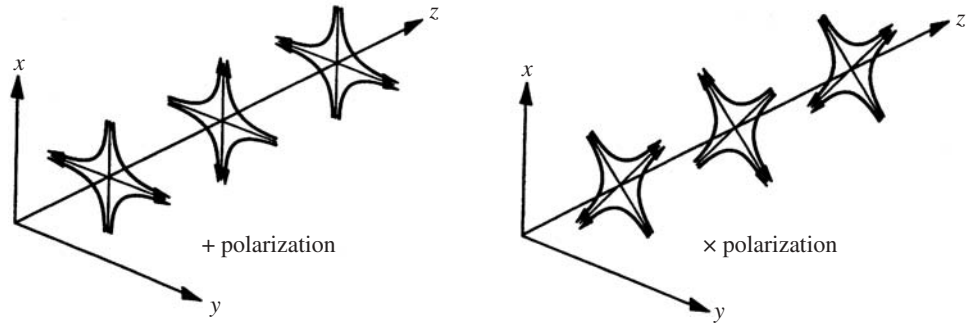


Figure 3.7 The lines of force associated with the two polarizations of a gravitational wave. Reprinted with permission of A. Abramovici *et al*. [14]. Copyright 1992 American Association for the Advancement of Science.

netic dipole radiation cannot produce any effect because of the conservation of momentum: any dipole radiation caused by the acceleration of an astronomical object is automatically cancelled by an equal and opposite change in momentum in nearby objects. Therefore, gravitational radiation is caused only by nonspherically symmetric accelerations of mass, which can be related to the quadrupole moment, and the oscillatory stretch and squeeze produced is then described by two dimensionless wave fields h_+ and $h_\times$, which are associated with the gravitational wave's two linear polarizations. If h_+ describes the amplitude of polarization with respect to the x- and y-axes in the horizontal plane, $h_\times$ describes the independent amplitude of polarization with respect to the rotated axes $x + y$ and $x - y$ (cf. Figure 3.7). The relative tidal effect a detector of length L may observe is then a linear combination of the two wave fields

$$\Delta L/L = a_+ h_+(t) + a_\times h_\times(t) \equiv h(t). \tag{3.32}$$

The proper derivation of the quadrupole formula for the energy loss rate through gravitational radiation of an oscillating body and the spatial strain $h(t)$ caused on bodies elsewhere cannot be carried out here, it requires general relativity to be carried out to high orders of covariant derivation. This complication is a benefit, however, because it renders the detection of gravitational radiation an extremely sensitive test of general relativity.

In a Newtonian approximation the strength of the waves from a nonspherical body of mass M, oscillating size $L(t)$, and quadrupole moment $Q(t) \approx ML^2$ at a distance r from Earth is

$$h(t) \approx \frac{G}{c^4 r}\frac{\mathrm{d}^2 Q(t)}{\mathrm{d}t^2} = \frac{G}{c^4 r} 2Mv(t)^2 = \frac{4G}{c^4 r}E(t), \tag{3.33}$$

where G is the Newtonian constant, v is the internal velocity, and $E = \frac{1}{2}Mv^2$ is the nonspherical part of the internal kinetic energy. The factor c^4 is introduced only to make $h(t)$ dimensionless.

Sources of Gravitational Waves. From this formula one can work out that a nonspherically symmetric supernova collapse at the centre of our Galaxy will give

rise to waves of amplitude $h \approx 10^{-19}$ causing a subnuclear stretch and squeeze of an object 1 km in length by 10^{-16} m. A spherically symmetric supernova collapse causes no waves. In a catastrophic event such as the collision of two neutron stars or two stellar-mass black holes in which E/c^2 is of the order of one solar mass, Equation (3.33) gives $h \approx 10^{-20}$ at the 16 Mpc distance of the Virgo cluster of galaxies, and $h \approx 10^{-21}$ at a distance of approximately 200 Mpc.

The signals one can expect to observe in the amplitude range $h \approx 10^{-21}$–10^{-20} with the next generation of detectors are bursts due to the coalescence of neutron-star binaries during their final minutes and seconds (in the high frequency band 1–10^4 Hz), and periodic waves from slowly merging galactic binaries and extragalactic massive black hole binaries (low-frequency band 10^{-4}–10^{-2} Hz), which are stable over hundreds to millions of years. The timing of millisecond binary pulsars such as the PSR 1913 + 16 belong to the very low-frequency band of 10^{-9}–10^{-7} Hz. In this band, processes in the very early Universe may also act as sources.

Merger waves from superheavy black holes with $10^6 M_\odot$ mass may be so strong that both their direction and their amplitude can be determined by monitoring the waves while the detector rotates around the Sun. This may permit researchers to identify the source with the parallax method and to determine the distance to it with high precision. Combined with redshift measurements of the source, one could determine not only H_0 but even the deceleration parameter q_0 of the Universe. Thus the detection of gravitational waves from black holes would go beyond testing general relativity to determining fundamental cosmological parameters of the Universe.

The dynamics of a hole–hole binary can be divided into three epochs: inspiral, merger and ringdown. The inspiral epoch ends when the holes reach their last stable orbit and begin plunging toward each other. Then the merger epoch commences, during which the binary is a single nonspherical black hole undergoing highly nonlinear space-time oscillations and vibrations of large amplitude. In the ringdown epoch, the oscillations decay due to gravitational wave emission, leaving finally a spherical, spinning black hole.

Gravitational Wave Detection. Detection with huge metal bars as resonant antennas was started by *Joseph Weber* in 1969. These detectors couple to one axis of the eigenmodes of the incoming wave, and one then expects to observe a change in the state of oscillation. Today several coordinated and aligned cryogenic bar detectors are in coordinated operation with sensitivities of approximately 10^{-21} Hz^{-1}. The detectors are tuned to see approximately 1 ms bursts occurring within a bandwidth of the order of 1 Hz. In order to eliminate random noise, the data from several detectors are analysed for coincidences.

To improve the signal-to-noise ratio in the high-frequency range one turns to Michelson interferometers with very long arms. The principle is illustrated in Figure 3.8. A laser beam is split, travels in two orthogonal directions to mirrors, and returns to be recombined and detected. A gravitational wave with either the h_+ or $h_\times$ component coinciding with the interferometer axes would lengthen the round-trip distance to one mirror and shorten it to the other. This would be

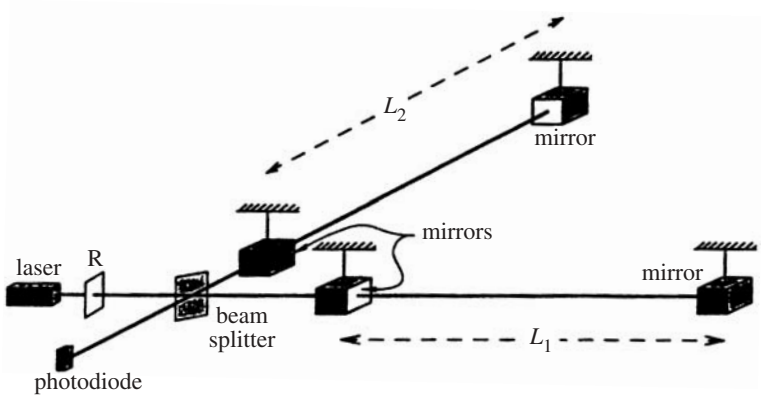


Figure 3.8 A schematic view of a LIGO-type interferometer. Reprinted with permission of A. Abramovici *et al*. [14]. Copyright 1992 American Association for the Advancement of Science.

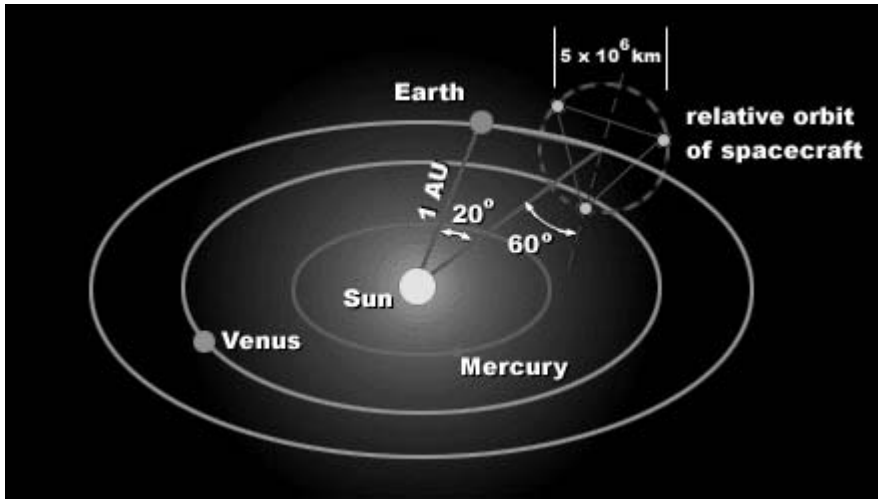


Figure 3.9 LISA orbit and position in the Solar System [16].

observable as a mismatch of waves upon recombination, and hence as a decrease in the observed combined intensity of the laser. For isolation against mechanical disturbances the optical components are carefully suspended in vacuum. The arm lengths in present detectors range from 300 m (TAMA in Japan) and 600 m (GEO600 in Germany) to 3 km (VIRGO in Italy) and 4 km (LIGO at two locations in the US). Sensitivities of 10^{-21}–10^{-22} Hz^{-1} can be reached in the high-frequency range. The range is limited to less than approximately 10^4 Hz by photo-electron shot noise in the components of the interferometer.

To study sources in the low-frequency range one has to put the interferometer into space orbiting Earth. This is necessary in order to avoid low-frequency seismic noise on the ground and thermally induced medium-frequency motion in the

atmosphere. The spectacular solution is the detector LISA (Laser Interferometer Space Antenna) consisting of three identical spacecraft, forming an equilateral triangle in space, with sidelength 5 million km, trailing Earth by 20° in a heliocentric orbit (cf. Figure 3.9). From each spacecraft a 1 W beam is sent to the two other remote spacecrafts via a telescope, is reflected by a platinum–gold cubic test mass, and the same telescopes are then used to focus the very weak returning beams. The interference signals from each arm are combined by on-board computers to perform the multiple-arm interferometry required to cancel the phase-noise common to all three arms. Fluctuations in the optical paths between the test masses can be measured to sub-angstrom precision, which, when combined with the large separation of the spacecraft, allows LISA to detect gravitational-wave strain down to a level of order 10^{-23} in one year of observation, with a signal-to-noise ratio of 5.

For a review of both experimental and theoretical aspects of all detectors currently operating or planned, see [15]. Since LISA is only scheduled to be launched in 2011, the interested reader is recommended to follow the ESA and NASA homepages [16].

Problems

1. Calculate the gravitational redshift in wavelength for the 769.9 nm potassium line emitted from the Sun's surface [1].

2. Derive the deflection angle (3.4) using Equations (1.26) and (3.5).

3. Derive Equation (3.12). What are the amplification of the individual images [4]?

4. A galaxy at $z = 0.9$ contains a quasar showing redshift $z = 1.0$. Supposing that this additional redshift of the quasar is caused by its proximity to a black hole, how many Schwarzschild radii away from the black hole does the light emitted by the quasar originate?

5. Estimate the gravitational redshift z of light escaping from a galaxy of mass $10^9 M_\odot$ after being emitted from a nearby star at a radial distance of 1 kpc from the centre of the galaxy. (Assume that all matter in the galaxy is contained within that distance [17].)

6. Light is emitted horizontally *in vacuo* near the Earth's surface, and falls freely under the action of gravity. Through what vertical distances has it fallen after travelling 1 km? Calculate the radial coordinate (expressed in Schwarzschild radii) at which light travels in a circular path around a body of mass M [17].

7. Derive Equation (3.25).

Chapter Bibliography

[1] Kenyon, I. R. 1990 *General relativity*. Oxford University Press, Oxford.
[2] Will, C. M. 1993 *Theory and experiment in gravitational physics*, revised edn. Cambridge University Press, Cambridge.
[3] See http://oposite.stsci.edu/pubinfo/pictures.html.
[4] Peacock, J. A. 1999 *Cosmological physics*. Cambridge University Press, Cambridge.
[5] Straumann, N. 2002 *Matter in the Universe*, Space Science Series of ISSI, vol. 14. Kluwer. (Reprinted from *Space Sci. Rev.* **100**, 29.)
[6] Hawking, S. and Penrose, R. 1996 *The nature of space and time*. Princeton University Press, Princeton, NJ.
[7] Bekenstein, J. 1973 *Phys. Rev.* D **7**, 2333.
[8] Hawking, S. W. 1974 *Nature* **248**, 30.
[9] Hawking, S. W. 1975 *Commun. Math. Phys.* **43**, 199.
[10] Schödel, R. *et al*. 2002 *Nature* **419**, 694.
[11] Gerssen, J. *et al*. 2002 *Astron. J.* **124**, 3270.
[12] Miyoshi, M. *et al*. 1995 *Nature* **373**, 127
[13] Wilkes, B. J. *et al*. 1995 *Astrophys. J.* **455**, L13.
[14] Abramovici, A. *et al*. 1992 *Science* **256**, 325.
[15] Maggiore, M. 2000 *Phys. Rep.* **331**, 283.
[16] See http://sci.esa.int/home/lisa/ and http://lisa.jpl.nasa.gov/.
[17] Berry, M. V. 1989 *Principles of cosmology and gravitation*. Adam Hilger, Bristol.

4

Cosmological Models

In Section 4.1 we turn to the 'concordance' or Friedmann–Lemaître–Robertson–Walker (FLRW) model of cosmology, really only a paradigm based on Friedmann's and Lemaître's equations and the Robertson–Walker metric, which takes both energy density and pressure to be functions of time in a Copernican universe. Among the solutions are the Einstein universe and the Einstein–de Sitter universe, now known to be wrong, as we shall see in Section 4.4, and the currently accepted Friedmann–Lemaître universe, which includes a positive cosmological constant.

In Section 4.2 we describe the de Sitter model, which does not apply to the Universe at large as we see it now, but which may have dominated the very early universe, and which may be the correct description for the future.

In Section 4.3 we study dark energy in the form of quintessence and other alternatives to the cosmological constant, which try to remove some of the problems of the Friedmann–Lemaître model.

In Section 4.4 we examine some of the classical tests of cosmological models, arguing that what is called a test is more often a case of parameter estimation. We also give the values of some of the parameters entering the FLRW model.

4.1 Friedmann–Lemaître Cosmologies

Let us now turn to our main subject, a model describing our homogeneous and isotropic Universe for which the Robertson–Walker metric (2.31) was derived. Recall that it could be written as a 4×4 tensor with nonvanishing components (2.32) on the diagonal only, and that it contained the curvature parameter k.

Introduction to Cosmology Third Edition by Matts Roos
© 2003 John Wiley & Sons, Ltd ISBN 0 470 84909 6 (cased) ISBN 0 470 84910 X (pbk)

Friedmann's Equations. The stress–energy tensor $T_{\mu\nu}$ entering on the right-hand side of Einstein's equations (2.96) was given by Equation (2.92) in its diagonal form. For a comoving observer with velocity four-vector $v = (c, 0, 0, 0)$, the time–time component T_{00} and the space–space component T_{11} are then

$$T_{00} = \rho c^2, \qquad T_{11} = \frac{pR^2}{1 - k\sigma^2}, \tag{4.1}$$

taking g_{00} and g_{11} from Equation (2.32). We will not need T_{22} or T_{33} because they just duplicate the results without adding new dynamical information. In what follows we shall denote mass density by ρ and energy density by ρc^2. Occasionally, we shall use $\rho_{\mathrm{m}} c^2$ to denote specifically the energy density in all kinds of matter: baryonic, leptonic and unspecified dark matter. Similarly we use $\rho_{\mathrm{r}} c^2$ or ε to specify the energy density in radiation.

On the left-hand side of Einstein's equations (2.96) we need G_{00} and G_{11} to equate with T_{00} and T_{11}, respectively. We have all the tools to do it: the metric components $g_{\mu\nu}$ are inserted into the expression (2.73) for the affine connection, and subsequently we calculate the components of the Riemann tensor from the expression (2.76) using the metric components and the affine connections. This lets us find the Ricci tensor components that we need, R_{00} and R_{11}, and the Ricci scalar from Equations (2.77) and (2.78), respectively. All this would require several pages to work out (see, for example, [1, 2]), so I only give the result:

$$G_{00} = 3(cR)^{-2}(\dot{R}^2 + kc^2), \tag{4.2}$$

$$G_{11} = -c^{-2}(2R\ddot{R} + \dot{R}^2 + kc^2)(1 - k\sigma^2)^{-1}. \tag{4.3}$$

Here R is the cosmic scale factor $R(t)$, not to be confused with the Ricci scalar R. Substituting Equations (4.1)–(4.3) into Einstein's equations (2.96) we obtain two distinct dynamical relations for $R(t)$:

$$\frac{\dot{R}^2 + kc^2}{R^2} = \frac{8\pi G}{3}\rho, \tag{4.4}$$

$$\frac{2\ddot{R}}{R} + \frac{\dot{R}^2 + kc^2}{R^2} = -\frac{8\pi G}{c^2}p. \tag{4.5}$$

These equations were derived in 1922 by Friedmann, seven years before Hubble's discovery, at a time when even Einstein did not believe in his own equations because they did not allow the Universe to be static. Friedmann's equations did not gain general recognition until after his death, when they were confirmed by an independent derivation (in 1927) by Georges Lemaître (1894–1966). For now they will constitute the tools for our further investigations.

The expansion (or contraction) of the Universe is inherent to Friedmann's equations. Equation (4.4) shows that the rate of expansion, $\dot{R}$, increases with the mass density ρ in the Universe, and Equation (4.5) shows that it may accelerate. Subtracting Equation (4.4) from Equation (4.5) we obtain

$$\frac{2\ddot{R}}{R} = -\frac{8\pi G}{3c^2}(\rho c^2 + 3p), \tag{4.6}$$

which shows that the acceleration decreases with increasing pressure and energy density, whether mass or radiation energy. Thus it is more appropriate to talk about the *deceleration* of the expansion.

At our present time t_0 when the mass density is ρ_0, the cosmic scale is R_0, the Hubble parameter is H_0 and the density parameter Ω_0 is given by Equation (1.35), Friedmann's equation (4.4) takes the form

$$\dot{R}_0^2 = \tfrac{8}{3}\pi G R_0^2 \rho_0 - kc^2 = H_0^2 R_0^2 \Omega_0 - kc^2, \tag{4.7}$$

which can be rearranged as

$$kc^2 = H_0^2 R_0^2 (\Omega_0 - 1). \tag{4.8}$$

It is interesting to note that this reduces to the Newtonian relation (1.35). Thus the relation between the Robertson–Walker curvature parameter k and the present density parameter Ω_0 emerges: to the k values $+1$, 0 and -1 correspond an overcritical density $\Omega_0 > 1$, a critical density $\Omega_0 = 1$ and an undercritical density $0 < \Omega_0 < 1$, respectively. The spatially flat case with $k = 0$ is called the *Einstein–de Sitter universe*.

General Solution. When we generalized from the present H_0 to the time-dependent Hubble parameter $H(t) = \dot{a}/a$ in Equation (2.46), this also implied that the critical density (1.31) and the density parameter (1.35) became functions of time:

$$\rho_c(t) = \frac{3}{8\pi G} H^2(t), \tag{4.9}$$

$$\Omega(t) = \rho(t)/\rho_c(t). \tag{4.10}$$

Correspondingly, Equation (4.8) can be generalized to

$$kc^2 = H^2 R^2 (\Omega - 1). \tag{4.11}$$

If $k \neq 0$, we can eliminate kc^2 between Equations (4.8) and (4.11) to obtain

$$H^2 a^2 (\Omega - 1) = H_0^2 (\Omega_0 - 1), \tag{4.12}$$

which we shall make use of later.

It is straightforward to derive a general expression for the solution of Friedmann's equation (4.4). Replacing $\dot{R}/R$ by $\dot{a}/a$, inserting kc^2 from Equation (4.8), and replacing $(8\pi G/3)\rho$ by $\Omega(a)H_0^2$, Equation (4.4) furnishes a solution for $H(a)$:

$$H(a) \equiv \dot{a}/a = H_0 \sqrt{(1 - \Omega_0) a^{-2} + \Omega(a)}. \tag{4.13}$$

Here we have left the a dependence of $\Omega(a)$ unspecified. As we shall see later, various types of energy densities with different a dependences contribute.

Equation (4.13) can be used to solve for the *lookback time* $t(z)/t_0$ or $t(a)/t_0$ (normalized to the age t_0) since a photon with redshift z was emitted by writing it as an integral equation:

$$\int_0^{t(a)} \mathrm{d}t = \int_1^a \frac{\mathrm{d}a}{aH(a)}. \tag{4.14}$$

The age of the Universe at a given redshift is then $1 - t(z)/t_0$. We shall specify this in more detail later.

Einstein Universe. Consider now the static universe cherished by Einstein. This is defined by $R(t)$ being a constant R_0 so that $\dot{R} = 0$ and $\ddot{R} = 0$ and the age of the Universe is infinite. Equations (4.4) and (4.5) then reduce to

$$\frac{kc^2}{R_0^2} = \frac{8\pi}{3} G\rho_0 = -\frac{8\pi}{c^2} Gp_0. \tag{4.15}$$

In order that the mass density ρ_0 be positive today, k must be $+1$. Note that this leads to the surprising result that the pressure of matter p_0 becomes negative!

Einstein corrected for this in 1917 by introducing a constant Lorentz-invariant term $\lambda g_{\mu\nu}$ into Equation (2.95), where the *cosmological constant* λ corresponds to a tiny correction to the geometry of the Universe. Equation (2.95) then becomes

$$G_{\mu\nu} = R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R - \lambda g_{\mu\nu}. \tag{4.16}$$

In contrast to the first two terms on the right-hand side, the $\lambda g_{\mu\nu}$ term does not vanish in the limit of flat space-time. With this addition, Friedmann's equations take the form

$$\frac{\dot{R}^2 + kc^2}{R^2} - \frac{\lambda}{3} = \frac{8\pi G}{3}\rho, \tag{4.17}$$

$$\frac{2\ddot{R}}{R} + \frac{\dot{R}^2 + kc^2}{R^2} - \lambda = -\frac{8\pi G}{c^2} p. \tag{4.18}$$

A positive value of λ curves space-time so as to counteract the attractive gravitation of matter. Einstein adjusted λ to give a static solution, which is called the *Einstein universe*.

The pressure of matter is certainly very small, otherwise one would observe the galaxies having random motion similar to that of molecules in a gas under pressure. Thus one can set $p = 0$ to a good approximation. In the static case when $R = R_0$, $\dot{R}_0 = 0$ and $\ddot{R}_0 = 0$, Equation (4.17) becomes

$$\frac{kc^2}{R_0^2} - \frac{\lambda}{3} = \frac{8\pi G}{3}\rho_0.$$

It follows from this that in a spatially flat Universe

$$\rho_\lambda = \frac{\lambda}{8\pi G} = -\rho_0. \tag{4.19}$$

But Einstein did not notice that the static solution is unstable: the smallest imbalance between λ and ρ would make $\ddot{R}$ nonzero, causing the Universe to accelerate into expansion or decelerate into contraction. This flaw was only noticed by Eddington in 1930, soon after Hubble's discovery, in 1929, of the expansion that caused Einstein to abandon his belief in a static universe and to withdraw the cosmological constant. This he called 'the greatest blunder of my lifetime'.

The Friedmann–Lemaître Universe. If the physics of the vacuum looks the same to any inertial observer, its contribution to the stress–energy tensor is the same as Einstein's cosmological constant λ, as was noted by Lemaître. The λ term in Equation (4.16) is a correction to the geometrical terms in $G_{\mu\nu}$, but the mathematical content of Equations (4.17) and (4.18) are not changed if the λ terms are moved to the right-hand side, where they appear as corrections to the stress–energy tensor $T_{\mu\nu}$. Then the physical interpretation is that of an ideal fluid with energy density $\rho_\lambda = \lambda/8\pi G$ and negative pressure $p_\lambda = -\rho_\lambda c^2$. When the cosmological constant is positive, the gravitational effect of this fluid is a cosmic repulsion counteracting the attractive gravitation of matter, whereas a negative λ corresponds to additional attractive gravitation.

The cosmology described by Equations (4.17) and (4.18) with a positive cosmological constant is called the *Friedmann–Lemaître universe*. Such a universe is now strongly supported by observations of a nonvanishing λ (as we shall see in Section 4.4), so the Einstein–de Sitter universe, which has $\lambda = 0$, is a dead end.

In a Friedmann–Lemaître universe the total density parameter is conveniently split into a matter term, a radiation term and a cosmological constant term,

$$\Omega_0 = \Omega_{\mathrm{m}} + \Omega_{\mathrm{r}} + \Omega_\lambda, \tag{4.20}$$

where Ω_{r} and Ω_λ are defined analogously to Equations (1.35) and (4.10) as

$$\Omega_{\mathrm{r}} = \frac{\rho_{\mathrm{r}}}{\rho_{\mathrm{c}}}, \qquad \Omega_\lambda = \frac{\lambda}{8\pi G\rho_{\mathrm{c}}} = \frac{\lambda}{3H_0^2}. \tag{4.21}$$

Ω_{m}, Ω_{r} and Ω_λ are important dynamical parameters characterizing the Universe. If there is a remainder $\Omega_k \equiv \Omega_0 - 1 \neq 0$, this is called the *vacuum-energy* term.

Using Equation (4.19) we can find the value of λ corresponding to the attractive gravitation of the present mass density:

$$-\lambda = 8\pi G\rho_0 = 3\Omega_0 H_0^2 \approx 1.3 \times 10^{-52} c^2\ \mathrm{m}^{-2}. \tag{4.22}$$

No quantity in physics this small has ever been known before. It is extremely uncomfortable that λ has to be fine-tuned to a value which differs from zero only in the 52nd decimal place(in units of $c = 1$). It would be much more natural if λ were exactly zero. This situation is one of the enigmas which will remain with us to the end of this book. As we shall see, a repulsive gravitation of this kind may have been of great importance during the first brief moments of the existence of the Universe, and it appears that the present Universe is again dominated by a global repulsion.

Energy-Momentum Conservation. Let us study the solutions of Friedmann's equations in the general case of nonvanishing pressure p. Differentiating Equation (4.4) with respect to time,

$$\frac{\mathrm{d}}{\mathrm{d}t}(\dot{R}^2 + kc^2) = \frac{8\pi G}{3}\frac{\mathrm{d}}{\mathrm{d}t}(\rho R^2),$$

we obtain an equation of second order in the time derivative:

$$2\dot{R}\ddot{R} = \tfrac{8}{3}\pi G(\dot{\rho}R^2 + 2\rho R\dot{R}). \tag{4.23}$$

Using Equation (4.6) to cancel the second-order time derivative and multiplying through by c^2/R^2, we obtain a new equation containing only first-order time derivatives:

$$\dot{\rho}c^2 + 3H(\rho c^2 + p) = 0. \tag{4.24}$$

This equation does not contain k and λ, but that is not a consequence of having started from Equations (4.4) and (4.5). If, instead, we had started from Equations (4.17) and (4.18), we would have obtained the same equation.

Note that all terms here have dimension energy density per time. In other words, Equation (4.24) states that the change of energy density per time is zero, so we can interpret it as the local *energy conservation law*. In a volume element $\mathrm{d}V$, $\rho c^2\,\mathrm{d}V$ represents the local decrease of gravitating energy due to the expansion, whereas $p\,\mathrm{d}V$ is the work done by the expansion. Energy does not have a global meaning in general relativity, whereas work does. If different forms of energy do not transform into one another, each form obeys Equation (4.24) separately.

As we have seen, Equation (4.24) follows directly from Friedmann's equations without any further assumptions. But it can also be derived in another way, perhaps more transparently. Let the total energy content in a comoving volume R^3 be

$$E = (\rho c^2 + p)R^3.$$

The expansion is *adiabatic* if there is no net inflow or outflow of energy so that

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}[(\rho c^2 + p)R^3] = 0. \tag{4.25}$$

If p does not vary with time, changes in ρ and R compensate and Equation (4.24) immediately follows.

The equation (4.24) can easily be integrated,

$$\int \frac{\dot{\rho}(t)c^2}{\rho(t)c^2 + p(t)}\,\mathrm{d}t = -3\int \frac{\dot{R}(t)}{R(t)}\,\mathrm{d}t = -3\int \frac{\dot{a}(t)}{a(t)}\,\mathrm{d}t, \tag{4.26}$$

if we know the relation between energy density and pressure—the *equation of state* of the Universe.

Entropy Conservation and the Equation of State. In contrast, the law of *conservation of entropy* S is not implied by Friedmann's equations, it has to be assumed specifically, as we shall demonstrate in Section 5.2,

$$\dot{S} = 0. \tag{4.27}$$

Then we can make an ansatz for the equation of state: let p be proportional to ρc^2 with some proportionality factor w which is a constant in time,

$$p = w\rho c^2. \tag{4.28}$$

In fact, one can show that this is the most general equation of state in a spacetime with Robertson–Walker metric. Inserting this ansatz into the integral in Equation (4.26) we find that the relation between energy density and scale is

$$\rho(a) \propto a^{-3(1+w)} = (1+z)^{3(1+w)}. \tag{4.29}$$

Here we use z as well as a because astronomers prefer z since it is an observable. In cosmology, however, it is better to use a or R for two reasons. Firstly, redshift is a property of light, but freely propagating light did not exist at times when $z \gtrsim 1000$, so z is then no longer a true observable. Secondly, it is possible to describe the future in terms of $a > 1$, but redshift is then not meaningful.

The value of the proportionality factor w in Equations (4.28) and (4.29) follows from the adiabaticity condition. Leaving the derivation of w for a more detailed discussion in Section 5.2, we shall anticipate here its value in three special cases of great importance.

(i) A *matter-dominated* universe filled with nonrelativistic cold matter in the form of pressureless nonradiating dust for which $p = 0$. From Equation (4.28) then, this corresponds to $w = 0$, and the density evolves according to

$$\rho_{\mathrm{m}}(a) \propto a^{-3} = (1+z)^3. \tag{4.30}$$

It follows that the evolution of the density parameter Ω_{m} is

$$\Omega_{\mathrm{m}}(a) = \Omega_{\mathrm{m}} \frac{H_0^2}{H^2} a^{-3}.$$

Solving for $H^2 a^2 \Omega$ and inserting it into Equation (4.13), one finds the evolution of the Hubble parameter:

$$H(a) = H_0 a^{-1} \sqrt{1 - \Omega_{\mathrm{m}} + \Omega_{\mathrm{m}} a^{-1}} = H_0 (1+z) \sqrt{1 + \Omega_{\mathrm{m}} z}. \tag{4.31}$$

(ii) A *radiation-dominated* universe filled with an ultra-relativistic hot gas composed of elastically scattering particles of energy density ε. Statistical mechanics then tells us that the equation of state is

$$p_{\mathrm{r}} = \tfrac{1}{3}\varepsilon = \tfrac{1}{3}\rho_{\mathrm{r}} c^2. \tag{4.32}$$

This evidently corresponds to $w = \frac{1}{3}$, so that the radiation density evolves according to

$$\rho_{\mathrm{r}}(a) \propto a^{-4} = (1+z)^4. \tag{4.33}$$

(iii) The *vacuum-energy* state corresponds to a flat, static universe ($\ddot{R} = 0$, $\dot{R} = 0$) without dust or radiation, but with a cosmological term. From Equations (4.17) and (4.18) we then obtain

$$p_\lambda = -\rho_\lambda c^2, \qquad w = -1. \tag{4.34}$$

Thus the pressure of the vacuum energy is negative, in agreement with the definition in Equation (4.19) of the vacuum-energy density as a negative quantity. In the equation of state (4.28), ρ_λ and p_λ are then scale-independent constants.

Early Time Dependence. It follows from the above scale dependences that the curvature term in Equation (4.17) obeys the following inequality in the limit of small R:

$$\frac{kc^2}{R^2} \ll \frac{8\pi G}{3}\rho + \frac{\lambda}{3}. \tag{4.35}$$

In fact, this inequality is always true when

$$k = +1, \qquad p > -\tfrac{1}{3}\rho c^2, \qquad w > -\tfrac{1}{3}, \qquad \lambda > 0. \tag{4.36}$$

Then we can neglect the curvature term and the λ term in Equation (4.17), which simplifies to

$$\frac{\dot{a}}{a} = H(t) = \left(\frac{8\pi G}{3}\rho\right)^{1/2} \propto a^{-3(1+w)/2}. \tag{4.37}$$

Let us now find the time dependence of a by integrating this differential equation:

$$\int \mathrm{d}a\, a^{-1+3(1+w)/2} \propto \int \mathrm{d}t,$$

to obtain the solutions

$$a^{3(1+w)/2} \propto t \quad \text{for } w \neq -1, \qquad \ln a \propto t \quad \text{for } w = -1.$$

Solving for a,

$$a(t) \propto t^{2/3(1+w)} \quad \text{for } w \neq -1, \qquad a(t) \propto \mathrm{e}^{\text{const.}\cdot t} \quad \text{for } w = -1. \tag{4.38}$$

In the two epochs of matter domination and radiation domination we know the value of w. Inserting this we obtain the time dependence of a for a matter-dominated universe,

$$a(t) \propto t^{2/3}, \tag{4.39}$$

and for a radiation-dominated universe,

$$a(t) \propto t^{1/2}. \tag{4.40}$$

Big Bang. We find the starting value of the scale of the Universe independently of the value of k in the curvature term neglected above:

$$\lim_{t\to 0} a(t) = 0. \tag{4.41}$$

In the same limit the rate of change $\dot{a}$ is obtained from Equation (4.37) with any w obeying $w > -1$:

$$\lim_{t\to 0} \dot{a}(t) = \lim_{t\to 0} a^{-1}(t) = \infty. \tag{4.42}$$

It follows from Equations (4.32) and (4.33) that an early radiation-dominated Universe was characterized by extreme density and pressure:

$$\lim_{t\to 0} \rho_{\mathrm{r}}(t) = \lim_{t\to 0} a^{-4}(t) = \infty,$$
$$\lim_{t\to 0} p_{\mathrm{r}}(t) = \lim_{t\to 0} a^{-4}(t) = \infty.$$

In fact, these limits also hold for any w obeying $w > -1$.

Actually, we do not even need an equation of state to arrive at these limits. Provided $\rho c^2 + 3p$ was always positive and λ negligible, we can see from Equations (4.6) and (4.18) that the Universe has always decelerated. It then follows that a must have been zero at some time in the past. Whether Friedmann's equations can in fact be trusted to that limit is another story which we shall come back to later. The time $t = 0$ was sarcastically called the *Big Bang* by Fred Hoyle, who did not like the idea of an expanding Universe starting from a singularity, but the name has stuck.

Late Einstein–de Sitter Evolution. The conclusions we derived from Equation (4.35) were true for past times in the limit of small a. However, the recent evolution and the future depend on the value of k and on the value of λ. For $k = 0$ and $k = -1$ the expansion always continues, following Equation (4.38), and a positive value of λ boosts the expansion further.

In a matter-dominated Einstein–de Sitter universe which is flat and has $\Omega_\lambda = 0$, Friedmann's equation (4.4) can be integrated to give

$$t(z) = \frac{2}{3H_0}(1+z)^{-3/2}, \tag{4.43}$$

and the present age of the Universe at $z = 0$ would be

$$t_0 = \frac{2}{3H_0}. \tag{4.44}$$

In that case the size of the Universe would be $ct_0 = 2h^{-1}$ Gpc. Inserting the value of H_0 used in Equation (1.21), $H_0 \approx 68$–75 km s^{-1} Mpc^{-1}, one finds

$$t_0 \approx 8.6\text{–}9.6 \text{ Gyr}. \tag{4.45}$$

This is in obvious conflict with t_0 as determined from the ages of the oldest known star in the Galaxy in Equation (1.23), 14.1 ± 2.5 Gyr. Thus the flat-universe model with $\Omega_\lambda = 0$ is in trouble.

Evolution of a Closed Universe. In a closed matter-dominated universe with $k = +1$ and $\lambda = 0$, the curvature term kc^2/R^2 drops with the second power of R, while, according to Equation (4.30), the density drops with the third power, so the inequality (4.35) is finally violated. This happens at a scale $R_{\max}$ such that

$$R_{\max}^{-2} = \frac{8\pi G\rho_{\rm m}}{3c^2}, \tag{4.46}$$

and the expansion halts because $\dot{R} = 0$ in Equation (4.4). Let us call this the *turnover time* $t_{\max}$. At later times the expansion turns into contraction, and the Universe returns to zero size at time $2t_{\max}$. That time is usually called the *Big Crunch*. For $k = +1$ Friedmann's equation (4.4) then takes the form

$$\frac{{\rm d}R}{{\rm d}t} = \sqrt{\frac{8\pi}{3}G\rho_{\rm m}(R)R^2 - c^2}.$$

Then $t_{\max}$ is obtained by integrating t from 0 to $t_{\max}$ and R from 0 to $R_{\max}$,

$$t_{\max} = \frac{1}{c}\int_0^{R_{\max}} \mathrm{d}R \left(\frac{8\pi G}{3c^2}\rho_{\mathrm{m}}(R)R^2 - 1\right)^{-1/2}. \tag{4.47}$$

To solve the R integral we need to know the energy density $\rho_{\mathrm{m}}(R)$ in terms of the scale factor, and we need to know $R_{\max}$. Let us take the mass of the Universe to be M. We have already found in Equation (2.42) that the volume of a closed universe with Robertson–Walker metric is

$$V = 2\pi^2 R^3.$$

Since the energy density in a matter-dominated universe is mostly pressureless dust,

$$\rho_{\mathrm{m}} = \frac{M}{V} = \frac{M}{2\pi^2 R^3}. \tag{4.48}$$

This agrees perfectly with the result (4.30) that the density is inversely proportional to R^3. Obviously, the missing proportionality factor in Equation (4.30) is then $M/2\pi^2$. Inserting the density (4.48) with $R = R_{\max}$ into Equation (4.46) we obtain

$$R_{\max} = \frac{4MG}{3\pi c^2}. \tag{4.49}$$

We can now complete the integral in Equation (4.47):

$$t_{\max} = \frac{\pi}{2c}R_{\max} = \frac{2MG}{3c^3}. \tag{4.50}$$

Although we might not know whether we live in a closed universe, we certainly know from the ongoing expansion that $t_{\max} > t_0$. Using the lower limit for t_0 from Equation (1.21) we find a lower limit to the mass of the Universe:

$$M > \frac{3t_0c^3}{2G} = 1.25 \times 10^{23} M_\odot. \tag{4.51}$$

Actually, the total mass inside the present horizon is estimated to be about $10^{22} M_\odot$.

The dependence of $t_{\max}$ on Ω_{m} can also be obtained:

$$t_{\max} = \frac{\pi\Omega_{\mathrm{m}}}{2H_0(\Omega_{\mathrm{m}} - 1)^{3/2}}. \tag{4.52}$$

The Radius of the Universe. The spatial curvature is given by the Ricci scalar R introduced in Equation (2.78), and it can be expressed in terms of Ω:

$$R = 6H^2(\Omega - 1). \tag{4.53}$$

Obviously, R vanishes in a flat universe, and it is only meaningful when it is non-negative, as in a closed universe. It is conventional to define a 'radius of curvature' that is also valid for open universes:

$$r_{\mathrm{U}} \equiv \sqrt{\frac{6}{R}} = \frac{1}{H\sqrt{|\Omega - 1|}}. \tag{4.54}$$

For a closed universe, r_{U} has the physical meaning of the radius of a sphere.

Another interesting quantity is the Schwarzschild radius of the Universe, $r_{\mathrm{c,U}}$. Combining Equations (4.50) and (3.19) we find

$$r_{\mathrm{c,U}} = 3ct_{\max} > 12 \text{ Gpc}. \tag{4.55}$$

Comparing this number with the much smaller Hubble radius $3h^{-1}$ Gpc in Equation (1.14) we might conclude that we live inside a black hole! However, the Schwarzschild metric is static whereas the Hubble radius recedes in expanding Friedmann models with superluminal velocity, as was seen in Equation (2.51), so it will catch up with $r_{\mathrm{c,U}}$ at some time. Actually, it makes more sense to describe the Big Bang singularity as a *white hole*, which is a time-reversed black hole. A white hole only emits and does not absorb. It has a horizon over which nothing gets in, but signals from inside do get out.

Evolution of Open, Closed and Flat Universes. The three cases $k = -1, 0, +1$ with $\lambda = 0$ are illustrated qualitatively in Figure 4.1. All models have to be consistent with the scale and rate of expansion today, R_0 and $\dot{R}_0$, at time t_0. Following the curves back in time one notices that they intersect the time axis at different times. Thus what may be called time $t = 0$ is more recent in a flat universe than in an open universe, and in a closed universe it is even more recent.

Late Friedmann–Lemaître Evolution. When $\lambda > 0$, the recent past and the future take an entirely different course (we do not consider the case $\lambda < 0$, which is of mathematical interest only). Since ρ_λ and Ω_λ are then scale-independent constants, they will start to dominate over the matter term and the radiation term when the expansion has reached a given scale. Friedmann's equation (4.18) can then be written

$$\frac{2\ddot{R}}{R} = 3H_0^2\Omega_\lambda.$$

From this one sees that the expansion will accelerate regardless of the value of k. In particular, a closed universe with $k = +1$ will ultimately not contract, but expand at an accelerating pace.

Let us now return to the general expression (4.14) for the normalized age $t(z)/t_0$ or $t(a)/t_0$ of a universe characterized by k and energy density components Ω_{m}, Ω_{r} and Ω_λ. Inserting the Ω components into Equations (4.13) and (4.14) we have

$$\frac{\dot{a}^2}{a^2} = H^2(t) = H_0^2[(1-\Omega_0)a^{-2} + \Omega_{\mathrm{m}}(a) + \Omega_{\mathrm{r}}(a) + \Omega_\lambda(a)],$$

or

$$t(z) = \frac{1}{H_0}\int_0^{1/(1+z)} \mathrm{d}a \left[(1-\Omega_0) + \Omega_{\mathrm{m}}a^{-1} + \Omega_{\mathrm{r}}a^{-2} + \Omega_\lambda a^2\right]^{-1/2}. \tag{4.56}$$

The integral can easily be carried out analytically when $\Omega_\lambda = 0$. But this is now of only academic interest, since we know today that $\Omega_\lambda \approx 0.7$ (as we shall see in

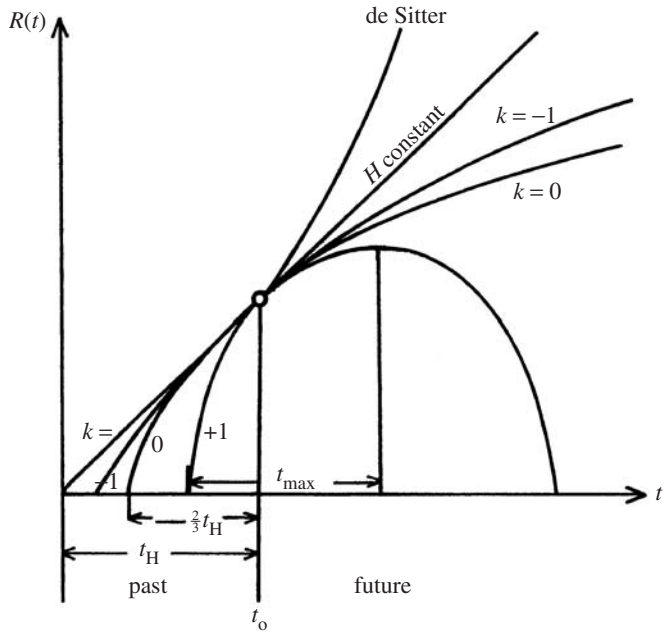


Figure 4.1 Time dependence of the cosmic scale $R(t)$ in various scenarios, all of which correspond to the same constant slope $H = H_0$ at the present time t_0. $k = +1$: a closed universe with a total lifetime $2t_{\max}$. It started more recently than a flat universe would have. $k = 0$: a flat universe which started $\frac{2}{3}t_{\rm H}$ ago. $k = -1$: an open universe which started at a time $\frac{2}{3}t_{\rm H} < t < t_{\rm H}$ before the present time. de Sitter: an exponential (inflationary) scenario corresponding to a large cosmological constant. This is also called the Lemaître cosmology.

Section 4.4). Thus the integral is best solved numerically (or analytically in terms of hypergeometric functions or the Weierstrass modular functions [3, 4]).

The lookback time is given by the same integral with the lower integration limit at $1/(1+z)$ and the upper limit at 1. The proper distance (2.38) is then

$$d_{\rm P}(z) = R_0\chi(z) = ct(z). \tag{4.57}$$

In Figure 4.2 we plot the lookback time $t(z)/t_0$ and the age of the Universe $1 - t(z)/t_0$ in units of t_0 as functions of redshift for the parameter values $\Omega_{\rm m} = 0.27$, $\Omega_\lambda = 1 - \Omega_{\rm m}$. At infinite redshift the lookback time is unity and the age of the Universe is zero.

Another important piece of information is that $\Omega_0 \approx 1.0$ (Table A.6). The vacuum term (4.8) (almost) vanishes, in which case we can conclude that the geometry of our Universe is (almost) flat. With $\Omega_0 = 1.0$ and $\Omega_{\rm r}$ well known (as we shall see in Section 5.2), the integral (4.56) really depends on only one unknown parameter, $\Omega_{\rm m} = 1 - \Omega_\lambda$.

From the values $\Omega_\lambda \approx 0.73$ and $\Omega_{\rm m} \approx 1 - 0.73 = 0.27$, one can conclude that the cosmological constant has already been dominating the expansion for some

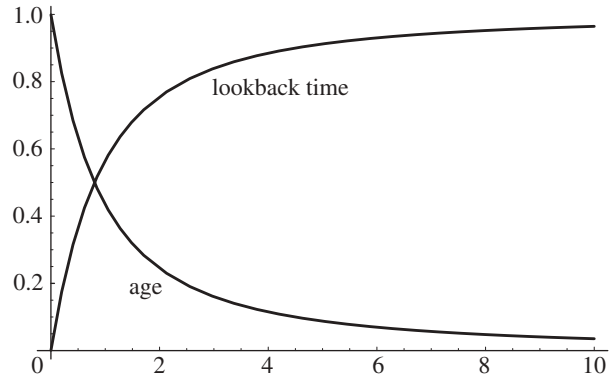


Figure 4.2 The lookback time and the age of the Universe normalized to t_0 as functions of redshift for the parameter values $\Omega_{\rm m} = 0.27$, $\Omega_\lambda = 1 - \Omega_{\rm m}$. For $z > 10$ see Figure 5.9.

time. The Universe began accelerating when Ω_λ and $\Omega_{\rm m}$ were equal, when

$$\frac{\Omega_{\rm m}(t)}{\Omega_\lambda} = \frac{0.27(1+z)^3}{0.73} = 1,$$

or at $z = 0.393$.

4.2 de Sitter Cosmology

Let us now turn to another special case for which Einstein's equations can be solved. Consider a homogeneous flat universe with the Robertson–Walker metric in which the density of pressureless dust is constant, $\rho(t) = \rho_0$. Friedmann's equation (4.17) for the rate of expansion including the cosmological constant then takes the form

$$\frac{\dot{a}(t)}{a(t)} = H, \tag{4.58}$$

where H is now a constant:

$$H = \sqrt{\frac{8\pi}{3} G\rho_0 + \frac{\lambda}{3}}. \tag{4.59}$$

This is clearly true when $k = 0$ but is even true for $k \neq 0$: since the density is constant and R increases without limit, the curvature term kc^2/R^2 will eventually be negligible. The solution to Equation (4.58) is obviously an exponentially expanding universe:

$$a(t) \propto \mathrm{e}^{Ht}. \tag{4.60}$$

This is drawn as the de Sitter curve in Figure 4.1. Substituting this function into the Robertson–Walker metric (2.31) we obtain the de Sitter metric

$$\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 - \mathrm{e}^{2Ht}(\mathrm{d}r^2 + r^2\,\mathrm{d}\theta^2 + r^2\sin^2\theta\,\mathrm{d}\phi^2) \tag{4.61}$$

with r replacing σ. In 1917 de Sitter published such a solution, setting $\rho = p = 0$, thus relating H directly to the cosmological constant λ. The same solution of

course follows even with $\lambda = 0$ if the density of dust ρ is constant. Eddington characterized the *de Sitter universe* as 'motion without matter', in contrast to the static *Einstein universe* that was 'matter without motion'.

If one introduces two test particles into this empty de Sitter universe, they will appear to recede from each other exponentially. The force driving the test particles apart is very strange. Let us suppose that they are at spatial distance rR from each other, and that λ is positive. Then the equation of relative motion of the test particles is given by Equation (4.5) including the λ term:

$$\frac{\mathrm{d}^2(rR)}{\mathrm{d}t^2} = \frac{\lambda}{3}rR - \frac{4\pi}{3}G(\rho + 3pc^{-2})rR. \tag{4.62}$$

The second term on the right-hand side is the decelerating force due to the ordinary gravitational interaction. The first term, however, is a force due to the vacuum-energy density, proportional to the distance r between the particles!

If λ is positive as in the Einstein universe, the force is repulsive, accelerating the expansion. If λ is negative, the force is attractive, decelerating the expansion just like ordinary gravitation. This is called an *anti-de Sitter* universe. Since λ is so small (cf. Equation (4.22)) this force will only be of importance to systems with mass densities of the order of the vacuum energy. The only known systems with such low densities are the large-scale structures, or the full horizon volume of cosmic size. This is the reason for the name *cosmological constant*. In Chapter 7 we shall meet inflationary universes with exponential expansion.

Although the world is not devoid of matter and the cosmological constant is small, the de Sitter universe may still be of more than academic interest in situations when ρ changes much more slowly than the scale R. The de Sitter metric then takes the form

$$\mathrm{d}s^2 = (1 - r^2H^2)\,\mathrm{d}t^2 - (1 - r^2H^2)^{-1}\,\mathrm{d}r^2 - r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2), \tag{4.63}$$

which resembles the Schwarzschild metric, Equation (3.21). There is an *inside* region in the de Sitter space at $r < H^{-1}$, for which the metric tensor component g_{00} is positive and g_{11} is negative. This resembles the region *outside* a black hole of Schwarzschild radius $r_\mathrm{c} = H^{-1}$, at $r > r_\mathrm{c}$, where g_{00} is positive and g_{11} is negative. Outside the radius $r = H^{-1}$ in de Sitter space and inside the Schwarzschild black hole these components of the metric tensor change sign. We shall come back to this metric in Chapter 10.

The interpretation of this geometry is that the de Sitter metric describes an expanding space-time surrounded by a black hole. Inside the region $r = H^{-1}$ no signal can be received from distances outside H^{-1} because there the metric corresponds to the inside of a black hole! In an anti-de Sitter universe the constant attraction ultimately dominates, so that the expansion turns into contraction. Thus de Sitter universes are open and anti-de Sitter universes are closed.

Let us study the particle horizon r_H in a de Sitter universe. Recall that this is defined as the location of the most distant visible object, and that the light from it started on its journey towards us at time t_H. From Equation (2.47) the particle horizon is at

$$r_\mathrm{H}(t) = R(t)\chi_\mathrm{ph} = R(t)\int_{t_\mathrm{H}}^{t_0}\frac{\mathrm{d}t'}{R(t')}. \tag{4.64}$$

Let us choose $t_{\rm H}$ as the origin of time, $t_{\rm H} = 0$. The distance $r_{\rm H}(t)$ as a function of the time of observation t then becomes

$$r_{\rm H}(t) = H^{-1}{\rm e}^{Ht}(1 - {\rm e}^{-Ht}). \tag{4.65}$$

The comoving distance to the particle horizon, $\chi_{\rm ph}$, quickly approaches the constant value H^{-1}. Thus for a comoving observer in this world the particle horizon would always be located at H^{-1}. Points which were inside this horizon at some time will be able to exchange signals, but events outside the horizon cannot influence anything inside this world.

The situation in a Friedmann universe is quite different. There the time dependence of a is not an exponential but a power of t, Equation (4.38), so that the comoving distance $\chi_{\rm ph}$ is an increasing function of the time of observation, not a constant. Thus points which were once at space-like distances, prohibited to exchange signals with each other, will be causally connected later, as one sees in Figure 2.1.

The importance of the de Sitter model will be illustrated later when we deal with exponential expansion at very early times in inflationary scenarios in Chapter 7.

4.3 Dark Energy

The introduction of the cosmological constant into our description of the Universe is problematic for at least three reasons. Firstly, as we noted in Equation (4.22), its present value is extremely small, in fact some 122 orders of magnitude smaller than theoretical expectations. The density is about

$$\rho_\lambda \approx 2.9 \times 10^{-47}\ {\rm GeV}^4.$$

If ρ_λ were even slightly larger, the repulsive force would cause the Universe to expand too fast so that there would not be enough time for the formation of galaxies or other gravitationally bound systems. This is called the *cosmological constant problem*.

Secondly, it raises the question of why the sum

$$\Omega_0 = \Omega_{\rm m} + \Omega_\lambda$$

is precisely 1.0 today when we are there to observe it, after an expansion of some 12 billion years when it was always greater than 1.0. The density of matter decreases like a^{-3}, while Ω_λ remains constant, so why has the cosmological constant been fine-tuned to come to dominate the sum only now? This is referred to as the *cosmic coincidence problem*.

Thirdly, we do not have the slightest idea what the λ energy consists of, only that it distorts the geometry of the Universe as if it were matter with strongly negative pressure, and acts as an *anti-gravitational* force which is unclustered at all scales. Since we know so little about it, we also cannot be sure that λ is constant in time, and that its equation of state is always $w_\lambda = -1$. When it is not constant it is often called *dark energy*.

Dark energy comes as a complete surprise. Nothing in big bang or inflationary cosmology predicted its existence. Therefore we also have no prediction for whether it is permanent, as a cosmological constant, or whether it will decay away in time.

Decaying Cosmological Constant. A dynamical approach to remove or alleviate the extreme need for fine-tuning λ is to choose it to be a slowly varying function of time, $\lambda(t)$. The initial conditions require $\lambda(t_{\text{Planck}}) \approx 10^{122}\lambda_0$, from which it decays to its present value at time t_0.

The Universe is then treated as a fluid composed of dust and dark energy in which the dark energy density, $\rho_\lambda(t) = \lambda(t)/8\pi G$, continuously transfers energy to the material component. Its equation of state is then of the form

$$p_\lambda = -\rho_\lambda\left(1 + \frac{1}{3}\frac{\mathrm{d}\ln\rho_\lambda(a)}{\mathrm{d}\ln a}\right). \tag{4.66}$$

In the classical limit when $\rho_\lambda(a)$ is a very slow function of a so that the derivative term can be ignored, one obtains the equation of state of the cosmological constant, $w_\lambda = -1$.

The advantage in removing the need for fine-tuning is, however, only replaced by another arbitrariness: an ansatz for $\lambda(t)$ is required and new parameters characterizing the timescale of the deflationary period and the transfer of energy from dark energy to dust must be introduced. Such phenomenological models have been presented in the literature [6, 7], and they can lead to testable predictions.

Scalar Fields. Instead of arguing about whether λ should be interpreted as a correction to the geometry or to the stress–energy tensor, we could go the whole way and postulate the existence of a new kind of energy, described by a slowly evolving scalar field $\varphi(t)$ that contributes to the total energy density together with the background (matter and radiation) energy density. This scalar field is assumed to interact only with gravity and with itself.

Since a scalar field is mathematically equivalent to a fluid with a time-dependent speed of sound, one can find potentials $V(\varphi)$ for which the dynamical vacuum acts like a fluid with negative pressure, and with an energy density behaving like a decaying cosmological constant. In comparison with plain $\lambda(t)$ models, scalar field cosmologies have one extra degree of freedom, since both a potential $V(\varphi)$ and a kinetic term $\frac{1}{2}\dot{\varphi}^2$ need to be determined.

The simplest equation of motion for a spatially homogeneous classical scalar field is the *Klein–Gordon* equation, which can be written

$$\ddot{\varphi} + 3H\dot{\varphi} + V'(\varphi) = 0, \tag{4.67}$$

where the prime indicates derivation with respect to φ. The energy density and pressure for a general scalar field enter in the diagonal elements of $T_{\mu\nu}$, and they are

$$\rho_\varphi c^2 = \tfrac{1}{2}\dot{\varphi}^2 + V(\varphi) \quad \text{and} \quad p_\varphi = \tfrac{1}{2}\dot{\varphi}^2 - V(\varphi), \tag{4.68}$$

respectively. Clearly the pressure is always negative if the evolution is so slow that the kinetic energy density $\frac{1}{2}\dot{\varphi}^2$ is less than the potential energy density. Note that in Equations (4.67) and (4.68) we have ignored terms describing spatial inhomogeneity which could also have been present.

The conservation of energy-momentum for the background component (matter and radiation, denoted by 'b') is Equation (4.24), which leads to the equation of state $w_{\rm b}$, and analogously one has for the scalar field

$$\dot{\rho}_\varphi c^2 + 3H(\rho_\varphi c^2 + p_\varphi) = 0$$

or

$$\dot{\rho}_\varphi + 3H\rho_\varphi(1 + w_\varphi) = 0. \tag{4.69}$$

As in Equation (4.29), the energy density of the scalar field decreases as $a^{-3(1+w_\varphi)}$. Inserting Equations (4.68) into Equation (4.69), one indeed obtains Equation (4.67). The equation of state of the φ field is then a function of the cosmological scale a (or time t or redshift z),

$$w_\varphi = \frac{\dot{\varphi}^2 + 2V(\varphi)}{\dot{\varphi}^2 - 2V(\varphi)}, \tag{4.70}$$

or, in some epochs, it can be a constant between 0 and -1.

However, dark energy defined this way and called *quintessence* turns out to be another *Deus ex machina* which not only depends on the parametrization of an arbitrary function $V(\varphi)$, but also has to be fine-tuned initially in a way similar to the cosmological constant.

Tracking Quintessence. In a somewhat less arbitrary model [8, 9], one constructs quintessence in such a way that its energy density is smaller than the background component for most of the history of the Universe, somehow tracking it with the same time dependence. As long as the field stays on the tracker solution and regardless of the value of w_φ in the radiation-domination epoch, w_φ automatically decreases to a negative value at time $t_{\rm eq}$ when the Universe transforms from radiation domination to matter domination. We saw in Equation (4.33) that radiation energy density evolves as a^{-4}—faster than matter energy density, a^{-3}. Consequently, $\rho_{\rm r}$ is now much smaller than $\rho_{\rm m}$.

But once w_φ is negative, ρ_φ decreases at a slower rate than $\rho_{\rm m}$ so that it eventually overtakes it. At that moment, $\varphi(t)$ slows to a near stop, causing w_φ to decrease toward -1, and tracking stops. Judging from the observed large value of the cosmological constant density parameter today, $\Omega_\lambda = 0.73$, this happened in the recent past when the redshift was $z \sim 2$–4. Quintessence is already dominating the total energy density, driving the Universe into a period of de Sitter-like accelerated expansion.

The tracker field should be an attractor in the sense that a very wide range of initial conditions for φ and $\dot{\varphi}$ rapidly approach a common evolutionary track, so that the cosmology is insensitive to the initial conditions. Thus the need for fine-tuning is entirely removed, the only arbitrariness remains in the choice of a

function $V(\varphi)$. With a judicious choice of parameters, the coincidence problem can also be considered solved, albeit by tuning the parameters ad hoc.

In Chapter 7 we shall come back to the inflationary de Sitter expansion following the Big Bang, which may also be caused by a scalar *inflaton field*. Here we just note that the initial conditions for the quintessence field can be chosen, if one so desires, to match the inflaton field.

Tracking behaviour with $w_\varphi < w_\mathrm{b}$ occurs [8, 9] for any potential obeying

$$\Gamma \equiv V''V/(V')^2 > 1, \tag{4.71}$$

and which is nearly constant over the range of plausible initial φ,

$$\frac{\mathrm{d}(\Gamma - 1)}{H\,\mathrm{d}t} \ll |\Gamma - 1|, \tag{4.72}$$

or if $-V'/V$ is a slowly decreasing function of φ. Many potentials satisfy these criteria, for instance power law, exponential times power law, hyperbolic, and Jacobian elliptic functions. For a potential of the generic form,

$$V(\varphi) = V_0(\varphi_0/\varphi)^{-\beta} \tag{4.73}$$

with β constant, one has a good example of a tracker field for which the kinetic and potential terms remain in a constant proportion.

The values of w_φ and Ω_φ depend both on $V(\varphi)$ and on the background. The effect of the background is through the $3H\dot{\varphi}$ term in the scalar field equation of motion (4.67): when w_b changes, H also changes, which, in turn, changes the rate at which the tracker field evolves down the potential.

The tracking potential is characterized as *slow rolling* when

$$\eta(\varphi) \equiv \frac{m^2_{\mathrm{Planck}}}{16\pi}\left(\frac{V''}{V}\right) \ll 1, \qquad \epsilon \equiv \frac{m^2_{\mathrm{Planck}}}{16\pi}\left(\frac{V'}{V}\right)^2 \ll 1, \tag{4.74}$$

meaning that $\ddot{\varphi}$ in Equation (4.67) and $\dot{\varphi}^2$ in Equation (4.68) are both negligible. At very early times, however, $-V'/V$ is slowly changing, but is itself not small. This establishes the important distinction between static and quasi-static quintessence with $w_\varphi \approx -1$ and dynamical quintessence with $w_\varphi > -1$. This means that the slow-roll approximation is not necessarily applicable to dynamical quintessence, and that the latter generally requires exact solution of the equation of motion (4.67) [10].

Given a potential like Equation (4.73) and fixing the current values of parameters Ω_m, Ω_r, Ω_φ, w_φ one can solve the equation of motion (4.67) by numerical integration. Finding the functions $\Gamma(a)$, $w_\varphi(\varphi)$, $w_\varphi(a)$ or $\Omega_\varphi(a)$ is a rather complicated exercise [8, 9, 10, 11]. In Figures 4.3 and 4.4 we show a few of these functions for inverse power potentials of the form (4.73). One can see that relatively fast-rolling dynamical quintessence also becomes static sooner or later, approaching $w_\varphi = -1$.

In this model the lookback time is given by

$$t(z) = \frac{1}{H_0}\int_1^{1/(1+z)} \mathrm{d}a[(1-\Omega_0) + \Omega_\mathrm{m}a^{-1} + \Omega_\mathrm{r}a^{-2} + \Omega_\varphi a^{2-3(1+w_\varphi)}]^{-1/2}. \tag{4.75}$$

For $1/(1+z) = 0$ this gives us the age (4.56) of the Universe, t_0.

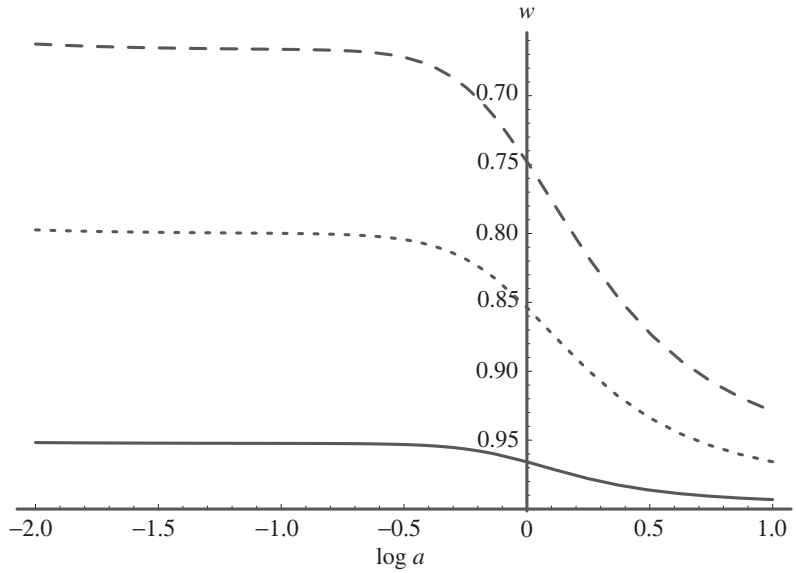


Figure 4.3 The quintessence equation of state $w_\varphi(a)$ for the inverse power potential (4.73) as a function of $\log a$ for $\beta = 0.1$ (solid line), $\beta = 0.5$ (dotted line), and $\beta = 1.0$ (dashed line). Here $\Omega_\varphi = \frac{2}{3}$, $\Omega_{\rm m} = \frac{1}{3}$.

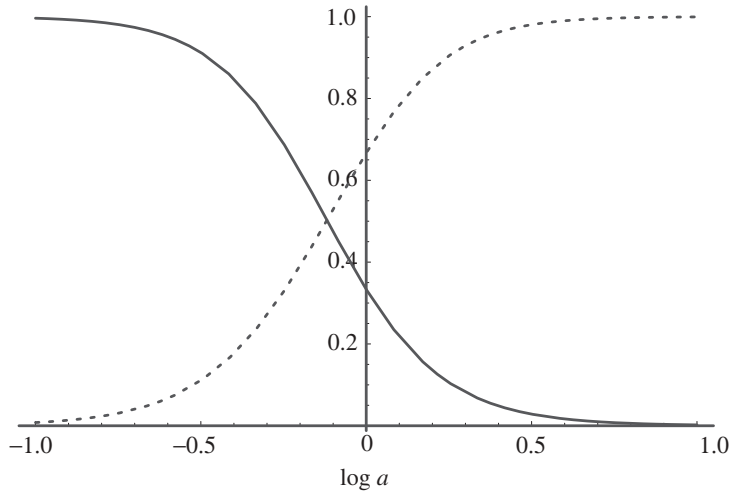


Figure 4.4 The quintessence density parameter $\Omega_\varphi(a)$ (solid line) and the background density parameter $\Omega_{\rm b}(a)$ (dotted line) for the inverse power potential (4.73) with $\beta = 0.5$ as a function of $\log a$. Note that $\log a = -1$ corresponds to $z = 9$.

Other Models. As already mentioned, the weakness of the tracking quintessence model is that the energy density for which the pressure becomes negative is set by an adjustable parameter which has to be fine-tuned to explain the cosmic coincidence. Surely one can do better by adding degrees of freedom, for instance by

letting $\varphi(t)$ interplay with the decaying cosmological constant $\lambda(t)$, or with the matter field, or with a second real scalar field $\psi(t)$ which dominates at a different time, or by taking $\varphi(t)$ to be complex, or by choosing a double-exponential potential, or by adding a new type of matter or a dissipative pressure to the background energy density. Surely the present acceleration could have been preceded by various periods of deceleration and acceleration. All of these alternatives have been proposed, but one generally then comes into the situation described by Wigner: 'with three parameters one can fit an elephant, with four one can make it wag its tail'.

One interesting alternative called *k-essence* [12, 13] comes at the cost of introducing a nonlinear kinetic energy density functional of the scalar field. The k-field tracks the radiation energy density until $t_{\rm eq}$, when a sharp transition from positive to negative pressure occurs, with $w_\varphi = -1$ as a consequence. The k-essence density $\rho_{\rm k}$ then drops below $\rho_{\rm m}$ and, thereafter, in the matter-dominated epoch the k-field does not track the background at all, it just stays constant. Thus the time of k-essence domination and accelerated expansion simply depends on $t_{\rm eq}$. However, this is another case of fine-tuning: $\rho_{\rm k}$ must drop precisely to the magnitude of the present-day ρ_λ.

Could the scalar field obey an equation of state with $w_\varphi < -1$? Such a situation would require rather drastic revisions of general relativity and would lead to infinite acceleration within a finite time [14]. Speculations have also appeared in the literature that the Universe might have undergone shorter periods of this type. It is well to remember that nothing is known about whether the cosmological constant is indeed constant or whether it will remain so, nor about the future behaviour of a quintessence field and its equation of state.

4.4 Model Testing and Parameter Estimation. I

In this chapter we have concentrated on the 'concordance' FLRW cosmological model with a nonvanishing cosmological constant in a spatially flat universe. But we should also give motivations for this choice and explain why other possibilities have lost their importance. The astronomical literature presently lists 13 tests of the standard model [5], not counting tests which may become important in the future. Some of them will be discussed in the present section, while others must be postponed until we have reached the necessary understanding of the underlying physics. We shall see that some of these classical tests do not really qualify as tests at all, and one of them, light deflection by lensing, is not a test of the cosmological model, but of general relativity. We already accounted for this and other tests of general relativity in Chapter 3.

Statistics. Let us take the meaning of the term 'test' from the statistical literature, where it is accurately defined [15]. When the hypothesis under test concerns the value of a parameter, the problems of *parameter estimation* and *hypothesis testing* are related; for instance, good techniques for estimation often lead to analogous

testing procedures. The two situations lead, however, to different conclusions, and should not be confused. If nothing is known *a priori* about the parameter involved, it is natural to use the data to estimate it. On the other hand, if a theoretical prediction has been made that the parameter should have a certain value, it may be more appropriate to formulate the problem as a test of whether the data are consistent with this value. In either case, the nature of the problem, estimation or test, must be clear from the beginning and consistent to the end. When two or more independent methods of parameter estimation are compared, one can talk about a *consistency test*.

A good example of this reasoning is offered by the discussion of Hubble's law in Section 1.4. Hubble's empirical discovery tested the *null hypothesis* that the Universe (out to the probed redshifts) expands. The test is a valid proof of the hypothesis for any value of H_0 that differs from zero at a chosen confidence level, CL%. Thus the value of H_0 is unimportant for the test, only its precision matters.

A determination of the value of H_0 is, however, not a test of a prediction, but a case of parameter estimation. The value of H_0 is then chosen to be at the maximum of the likelihood function or in the middle of a confidence range. For a Gaussian probability density function, a $\pm 1\sigma$ (one standard deviation) range represents a probability of 68.3%, a $\pm 2\sigma$ range represents a probability of 95.4%, and so on.

A combination of estimates such as those referred to in Section 1.4 furnishes a consistency test. The consistency is then quantified by the total sum of log-likelihood functions, which in the case of Gaussian probability density functions reduces to the well-known χ^2-test.

Expansion Time. The so-called timescale test compares the lookback time in Figure 4.2 at redshifts at which galaxies can be observed with t_0 obtained from other cosmochronometers inside our Galaxy, as discussed in Section 1.4. Thus we have to make do with a consistency test. At moderately high redshifts where the $\Omega_{\rm m}$ term dominates and Ω_λ can be neglected, Equation (4.56) can be written

$$H_0 t(z) \approx \frac{2}{3\sqrt{\Omega_{\rm m}}}(1+z)^{-3/2}. \tag{4.76}$$

Let us multiply the H_0 and t_0 values in Table A.2 to obtain a value for the dimensionless quantity

$$H_0 t_0 = 0.97 \pm 0.05. \tag{4.77}$$

As we already saw in Equation (4.44) this rules out the spatially flat matter-dominated Einstein–de Sitter universe in which $H_0 t_0 < \frac{2}{3}$.

Equations (4.76) and (4.77) can also be combined to give an estimate for $\Omega_{\rm m}$. Taking $t(3) = 2.4$ Gyr at $z = 3$, one finds

$$\Omega_{\rm m} = 0.23,$$

which is not very different from the figure that we quote in Table A.6.

The Magnitude–Redshift Relation. Equation (2.60) relates the apparent magnitude m of a bright source of absolute magnitude M at redshift z to the luminosity distance d_L. We noted in Section 1.4 that the peak brightness of SNe Ia can serve as remarkably precise standard candles visible from very far away; this determines M. Although the magnitude–redshift relation can be used in various contexts, we are only interested in testing cosmology.

The luminosity distance d_L is a function of z and the model-dependent dynamical parameters, primarily Ω_m, Ω_λ and H_0. In Section 1.4 and Figure 1.2 we already referred to measurements of H_0 based on HST supernova observations (as well as other types of measurement). Supernova observations furnish the best information on Ω_λ. The redshift can be measured in the usual way by observing the shift of spectral lines, but the supernova light-curve shape gives supplementary information: in the rest frame of the supernova the time dependence of light emission follows a standard curve, but a supernova at relativistic distances exhibits a broadened light curve due to time dilation.

Two groups [16, 17, 18, 19] have reported their analyses of 16 and 42 supernovae of type Ia, respectively, using somewhat different methods of light-curve fitting to determine the distance moduli, determining the parameters by maximum likelihood fits, and reaching the same conclusions. Following the analysis of Sullivan *et al*. [17, 18, 19], who define a 'Hubble-constant free' luminosity distance $D_L \equiv H_0 d_L$, the effective B-band (a standard blue filter) magnitude m_B becomes

$$m_B = M_B - 5\log H_0 + 25 + 5\log D_L(z,\ \Omega_m,\ \Omega_\lambda). \tag{4.78}$$

Figure 4.5 shows a Hubble plot [19] of m_B versus z for the supernovae studied. The solid curve represents an accelerating, spatially flat FLRW universe with

$$\Omega_\lambda = 1 - \Omega_m = 0.72. \tag{4.79}$$

Note that this value is not a test of the FLRW model since the model does not make a specific prediction for Ω_λ. The dotted curve in Figure 4.5 that is denoted 'EdS' represents the Einstein–de Sitter model, which predicts $\Omega_\lambda = 0$. Comparison of the curves then constitutes a test of the Einstein–de Sitter model (here the null hypothesis), with an overwhelming statistical significance [17, 18, 19].

Figure 4.6 shows the confidence regions of the fit in the $(\Omega_m, \Omega_\lambda)$-plane. One notices that the most precisely determined parameter is not Ω_λ or Ω_m, but the difference $\Omega_\lambda - \Omega_m$, which is measured along the flat-space line. The position of the best-fit-confidence region also shows that the supernova data allow a test of a decelerating versus an accelerating universe. A decelerated universe is disfavoured by a confidence of about 99.7% (3σ).

In general, the magnitude–redshift relation is a case of parameter estimation, and will remain so in the future when the study of more supernovae will allow more precise determination of Ω_λ and dependent quantities such as t_0, q_0, w_φ.

The Angular Diameter–Redshift Relation. In Equation (2.63) we related the angular size distance d_A to the proper distance $d_P(k, H_0, \Omega_m, \Omega_\lambda)$. In conventional

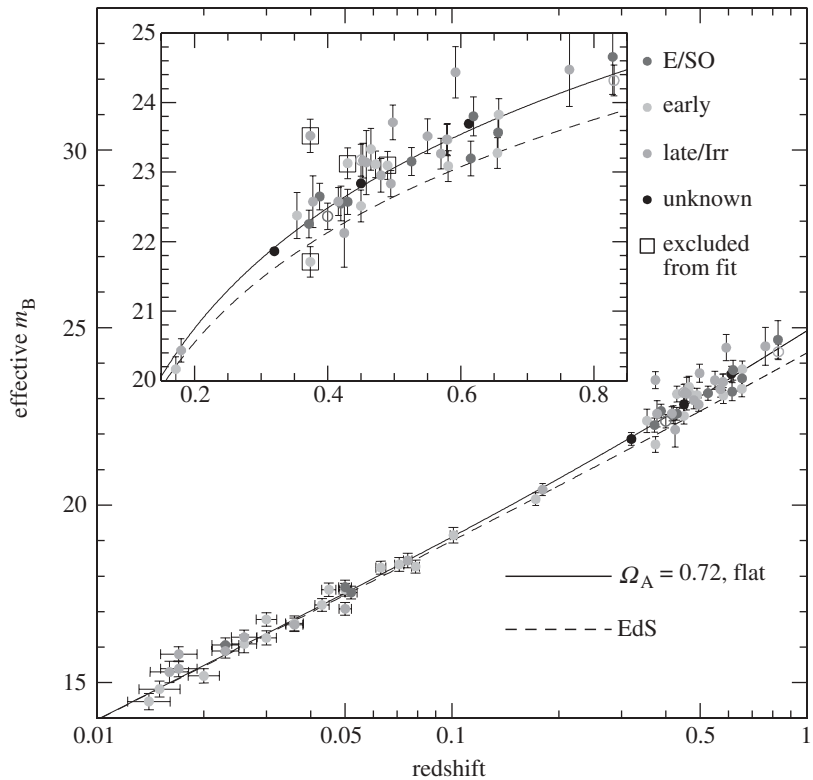


Figure 4.5 Hubble diagram of effective B-band magnitude versus redshift for the supernovae studied by Sullivan *et al*. [17, 18, 19]. The different round and boxed points correspond to different classes of host galaxy. Reproduced from M. Sullivan *et al*. [19], *The Hubble Diagram of type-Ia supernovae as a function of host galaxy morphology* by permission of Blackwell Publishing Ltd.

local physics with a single metric theory the relations (2.60) and (2.63) are physically equivalent. Thus our comments on to what extent Equation (2.63) and supernova data furnish tests or imply parameter estimation are the same as above.

Galaxy and Quasar Counts. The observable here is the number of galaxies or quasars within a comoving volume element. The difficulty with galaxies is that there are many more galaxies with low luminosities than with high luminosities, so this possible test depends on the understanding of the evolution of galaxy luminosities. But luminosity distances again depend on $H(z)$, which in turn depends on the unknown parameters in Equation (4.56). Thus galaxy counts do not appear to constitute a test, rather a method of parameter estimation in the same sense as the previous cases.

Quasars derive their luminosity from rotating accretion discs spiralling into massive black holes at the centres of galaxies. If one compares the number of lensed quasars seen with the number predicted by observations of intervening

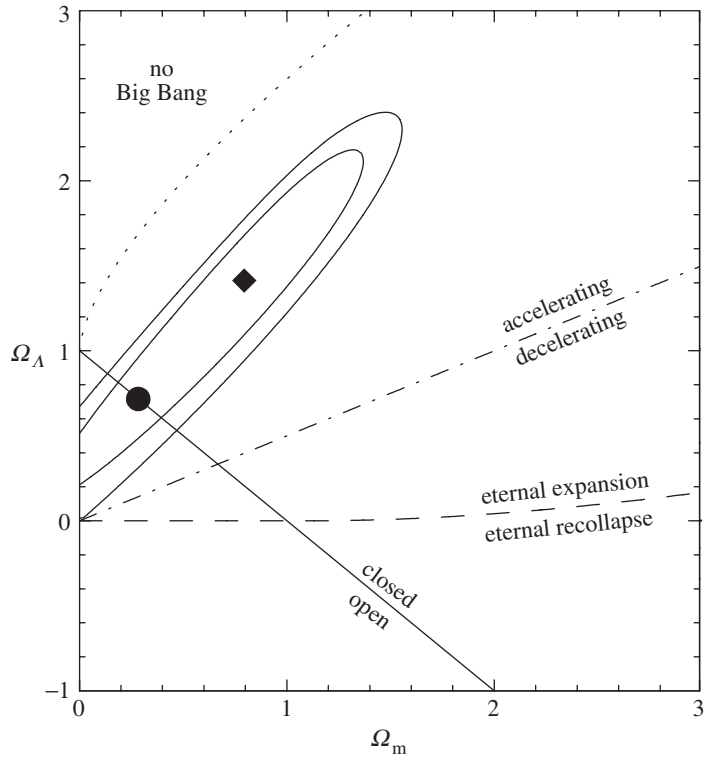


Figure 4.6 Confidence regions in $(\Omega_m, \Omega_\lambda)$ for the fitting procedures in Sullivan *et al.* [17, 18, 19]. The ellipses correspond to 68% and 90% confidence regions. The best-fitting general FLRW cosmology is denoted by a filled diamond, and the best-fitting flat cosmology by a filled circle. The solid line is the flat-space boundary between closed and open cosmologies, the dotted line is the boundary between finite and infinite t_0 ('no Big Bang'), the dashed line is the infinite expansion boundary, and the dot-dashed line separates accelerating and decelerating universes. (By courtesy of the Supernova Cosmology Project team.)

galaxies capable of lensing, one finds that the number of lensing events is about double the prediction, unless dark energy is present. This gives an independent estimate of the fraction of dark energy, about $\frac{2}{3}$, in agreement with Equation (4.79).

Problems

1. On the solar surface the acceleration caused by the repulsion of a nonvanishing cosmological constant λ must be much inferior to the Newtonian attraction. Derive a limiting value of λ from this condition.

2. In Newtonian mechanics, the cosmological constant λ can be incorporated by adding to gravity an outward radial force on a body of mass m, a distance r from the origin, of $F = +m\lambda r/6$. Assuming that $\lambda = -10^{-20}$ yr^{-2}, and that

F is the only force acting, estimate the maximum speed a body will attain if its orbit is comparable in size with the Solar System (0.5 light day) [20].

3. Einstein's static universe has $a \propto e^{Ht}$, zero curvature of its comoving coordinates ($k = 0$), and a proper density of all objects that is constant in time. Show that the comoving volume out to redshift z is $V(z) = \frac{4}{3}\pi(cz/H)^3$, and hence that the number-count slope for objects at typical redshift z becomes $[(3 + \alpha)\ln z]^{-1}$ for $z \gg 1$, where α is the spectral index for the objects [21].
4. Starting from Equation (4.56) with the parameters $\Omega_0 = 1$, $\Omega_r = 0$, show that the age of the Universe can be written

$$t_0 = \frac{2}{3H_0}\frac{\tanh^{-1}\sqrt{\Omega_\lambda}}{\sqrt{\Omega_\lambda}}.$$

5. Suppose that dark energy is described by an equation of state $w = -0.9$ which is constant in time. At what redshift did this dark energy density start to dominate over matter density? What was the radiation density at that time?

Chapter Bibliography

[1] Rich, J. 2001 *Fundamentals of cosmology*. Springer.
[2] Solà, J. 2001 *Nucl. Phys.* B **95**, 29.
[3] Cappi, A. 2001 *Astrophys. Lett. Commun.* **40**, 161.
[4] Kraniotis, G. V. and Whitehouse, S. B. 2002 *Classical Quantum Gravity* **19**, 5073.
[5] Peebles, P. J. E. and Ratra, B. 2003 *Rev. Mod. Phys.* **75**, 559.
[6] Lima, J. A. S. and Trodden, M. 1996 *Phys. Rev.* D **53**, 4280.
[7] Cunha, J. V., Lima, J. A. S. and Pires, N. 2002 *Astron. Astrophys.* **390**, 809.
[8] Steinhardt, P. J., Wang, L. and Zlatev, I. 1999 *Phys. Rev.* D **59**, 123504.
[9] Zlatev, I., Wang, L. and Steinhardt, P. J. 1999 *Phys. Rev. Lett.* **82**, 896.
[10] Bludman, S. A. and Roos, M. 2002 *Phys. Rev.* D **65**, 043503.
[11] Ng, S. C. C., Nunes, N. J. and Rosati, F. 2001 *Phys. Rev.* D **64**, 083510.
[12] Armendariz-Picon, C., Mukhanov, V. and Steinhardt, P. J. 2000 *Phys. Rev. Lett.* **85**, 4438.
[13] Armendariz-Picon, C., Mukhanov, V. and Steinhardt, P. J. 2001 *Phys. Rev.* D **63**, 103510.
[14] Caldwell, R. R., Kamionkowski, M. and Weinberg, N. V. 2003 *Phys. Rev. Lett.*. (In press).
[15] Eadie, W. T., Drijard, D., James, F. E., Roos, M. and Sadoulet, B. 1971 *Statistical methods in experimental physics*. North-Holland, Amsterdam.
[16] Riess, A. G. *et al*. 1998 *Astronom. J.* **116**, 1009.
[17] Perlmutter, S. *et al*. 1998 *Nature* **391**, 51.
[18] Perlmutter, S. *et al*. 1999 *Astrophys. J.* **517**, 565.
[19] Sullivan, M. *et al*. 2003 *Mon. Not. R. Astron. Soc.* **340**, 1057.
[20] Berry, M. V. 1989 *Principles of cosmology and gravitation*. Adam Hilger, Bristol.
[21] Peacock, J. A. 1999 *Cosmological physics*. Cambridge University Press, Cambridge.

5

Thermal History of the Universe

The Big Bang models describe the evolution of our Universe from a state of extreme pressure and energy density, when it was very much smaller than it is now. Matter as we know it does not stand up to extreme temperatures. The Sun is a plasma of ionized hydrogen, helium and other elements, but we know also that the stability of nuclei cannot withstand temperatures corresponding to a few MeV of energy. They decompose into elementary particles, which at yet higher temperatures decompose into even more elementary constituents under conditions resembling those met in high-energy particle colliders. An understanding of cosmology therefore requires that we study the laws and phenomena of very high-temperature plasmas during the early radiation era.

Motion of particles under electromagnetic interaction is described by the Maxwell–Lorentz equations. The motion of a particle in a central field of force F, as for instance an electron of charge e moving at a distance r around an almost static proton, is approximated well by the *Coulomb force*

$$F = \frac{e^2}{r^2}. \tag{5.1}$$

Note that this has the same form as Newton's law of gravitation, Equation (1.28). In the electromagnetic case the strength of the interaction is e^2, whereas the strength of the gravitational interaction is GMm_{G}. These two *coupling constants* are expressed in completely different units because they apply to systems of completely different sizes. For the physics of radiation, the gravitational interaction can be completely neglected but, for the dynamics of the expansion of the Universe, only the gravitational interaction is important because celestial objects are electrically neutral.

Introduction to Cosmology Third Edition by Matts Roos
© 2003 John Wiley & Sons, Ltd ISBN 0 470 84909 6 (cased) ISBN 0 470 84910 X (pbk)

In Section 5.1 we begin with the physics of photons and Planck's radiation law, which describes how the energy is distributed in an ensemble of photons in thermal equilibrium, the blackbody spectrum. We also introduce the properties of polarization and spin.

In Section 5.2 we introduce the important concept of entropy and we note that a universe filled with particles and radiation in thermal equilibrium must indeed have been radiation dominated at an early epoch. Comparing a radiation-dominated universe with one dominated by nonrelativistic matter in adiabatic expansion, we find that the relation between temperature and scale is different in the two cases. This leads to the conclusion that the Universe will not end in thermal death, as feared in the 19th century.

In Section 5.3 we meet new particles and antiparticles, fermions and bosons, some of their properties such as conserved quantum numbers, spin, degrees of freedom and energy spectrum, and a fair number of particle reactions describing their electroweak interactions in the primordial plasma.

In Section 5.4 we trace the thermal history of the Universe, starting at a time when the temperature was 10^{13} K. The Friedmann equations offer us the means of time-keeping as a function of temperature.

In Section 5.5 we continue the thermal history of photons and leptons from neutrino decoupling to electron decoupling to the cold microwave radiation of today.

In Section 5.6 we follow the thermal history of the nucleons for the momentous process of Big Bang nucleosynthesis (BBN) which has left us very important clues in the form of relic abundances of helium and other light nuclei. The nucleosynthesis is really a very narrow bottleneck for all cosmological models, and one which has amply confirmed the standard Big Bang model. We find that the baryonic matter present since nucleosynthesis is completely insufficient to close the Universe.

5.1 Photons

Electromagnetic radiation in the form of radio waves, microwaves, light, X-rays or γ-rays has a dual description: either as waves characterized by the wavelength λ and frequency $\nu = c/\lambda$, or as energy quanta, *photons*, γ. In the early days of quantum theory the wave–particle duality was seen as a logical paradox. It is now understood that the two descriptions are complementary, the wave picture being more useful to describe, for instance, interference phenomena, whereas the particle picture is needed to describe the kinematics of particle reactions or, for instance, the functioning of a photocell (this is what Einstein received the Nobel prize for!). Energy is not a continuous variable, but it comes in discrete packages: it is *quantized*. The quantum carried by an individual photon is

$$E = h\nu, \tag{5.2}$$

where h is Planck's constant. The wavelength and energy ranges corresponding to the different types of radiation are given in Table A.3.

Blackbody Spectrum. Let us study the thermal history of the Universe in the Big Bang model. At the very beginning the Universe was in a state of extreme heat and pressure, occupying an exceedingly small volume. Before the onset of the present epoch, in which most of the energy exists in the form of fairly cold matter, there was an era when the pressure of radiation was an important component of the energy density of the Universe, the era of *radiation domination*. As the Universe cooled, matter condensed from a hot plasma of particles and electromagnetic radiation, later to form material structures in the forms of clusters, galaxies and stars.

During that era no atoms or atomic nuclei had yet been formed, because the temperature was too high. Only the particles which later combined into atoms existed. These were the free electrons, protons, neutrons and various unstable particles, as well as their antiparticles. Their speeds were relativistic, they were incessantly colliding and exchanging energy and momentum with each other and with the radiation photons. A few collisions were sufficient to distribute the available energy evenly among them. On average they would then have the same energy, but some particles would have less than average and some more than average. When the collisions resulted in a stable energy spectrum, *thermal equilibrium* was established and the photons had the *blackbody spectrum* derived in 1900 by Max Planck.

Let the number of photons of energy $h\nu$ per unit volume and frequency interval be $n_\gamma(\nu)$. Then the photon number density in the frequency interval $(\nu, \nu + \mathrm{d}\nu)$ is

$$n_\gamma(\nu)\,\mathrm{d}\nu = \frac{8\pi}{c^3}\frac{\nu^2\,\mathrm{d}\nu}{\mathrm{e}^{h\nu/kT} - 1}. \tag{5.3}$$

At the end of the 19th century some 40 years was spent trying to find this formula using trial and error. With the benefit of hindsight, the derivation is straightforward, based on classical thermodynamics as well as on quantum mechanics, unknown at Planck's time.

Note that Planck's formula depends on only one parameter, the temperature T. Thus the energy spectrum of photons in thermal equilibrium is completely characterized by its temperature T. The distribution (5.3) peaks at the frequency

$$\nu_{\max} \simeq 10^{10} T \tag{5.4}$$

in units of hertz or cycles per second, where T is given in kelvin.

The total number of photons per unit volume, or the *number density* N_γ, is found by integrating this spectrum over all frequencies:

$$N_\gamma = \int_0^\infty n_\gamma(\nu)\,\mathrm{d}\nu \simeq 1.202\frac{2}{\pi^2}\left(\frac{kT}{c\hbar}\right)^3. \tag{5.5}$$

Here $\hbar$ represents the reduced Planck's constant $\hbar = h/2\pi$. The solution of the integral in this equation can be given in terms of Riemann's zeta-function; $\zeta(3) \approx 1.2020$.

Since each photon of frequency ν is a quantum of energy $h\nu$ (this is the interpretation Planck was led to, much to his own dismay, because it was in obvious conflict with classical ideas of energy as a continuously distributed quantity), the

total energy density of radiation is given by the *Stefan–Boltzmann law* after *Josef Stefan* (1835–1893) and *Ludwig Boltzmann* (1844–1906),

$$\varepsilon_{\rm r} = \int_0^\infty h\nu n_\nu(\nu)\,{\rm d}\nu = \frac{\pi^2}{15}\frac{k^4T^4}{\hbar^3c^3} \equiv a_{\rm S}T^4, \tag{5.6}$$

where all the constants are lumped into Stefan's constant

$$a_{\rm S} = 4723\ {\rm eV\,m^{-3}\,K^{-4}}.$$

A blackbody spectrum is shown in Figure 8.1.

Polarization. Consider a plane wave of monochromatic light with frequency ν moving along the momentum vector in the z direction. The components of the wave's electric field vector $\boldsymbol{E}$ in the (x, y)-plane oscillate with time t in such a way that they can be written

$$E_x(t) = a_x(t)\cos[\nu t - \theta_x(t)], \qquad E_y(t) = a_y(t)\cos[\nu t - \theta_y(t)], \tag{5.7}$$

where $a_x(t)$ and $a_y(t)$ are the amplitudes, and $\theta_x(t)$ and $\theta_y(t)$ are the phase angles.

A well-known property of light is its two states of *polarization*. Unpolarized light passing through a pair of polarizing sunglasses becomes vertically polarized. Unpolarized light reflected from a wet street becomes horizontally polarized. The advantage of polarizing sunglasses is that they block horizontally polarized light completely, letting all the vertically polarized light through. Their effect on a beam of unpolarized sunlight is to let, on average, every second photon through vertically polarized, and to block every other photon as if it were horizontally polarized: it is absorbed in the glass. Thus the intensity of light is also reduced to one-half.

Polarized and unpolarized light (or other electromagnetic radiation) can be described by the *Stokes parameters*, which are the time averages (over times much longer than $1/\nu$)

$$\left.\begin{aligned} I &\equiv \langle a_x^2\rangle + \langle a_y^2\rangle, & Q &\equiv \langle a_x^2\rangle - \langle a_y^2\rangle, \\ U &\equiv \langle 2a_xa_y\cos(\theta_x-\theta_y)\rangle, & V &\equiv \langle 2a_xa_y\sin(\theta_x-\theta_y)\rangle. \end{aligned}\right\} \tag{5.8}$$

The parameter I gives the intensity of light, which is always positive definite. The electromagnetic field is unpolarized if the two components in Equation (5.7) are uncorrelated, which translates into the condition $Q = U = V = 0$. If two components in Equation (5.7) are correlated, they either describe light that is *linearly polarized* along one direction in the (x, y)-plane, or *circularly polarized* in the plane. In the linear case $U = 0$ or $V = 0$, or both. Under a rotation of angle ϕ in the (x, y)-plane, the quantity $Q^2 + U^2$ is an invariant (Problem 2) and the *orientation* of the polarization

$$\alpha \equiv \tfrac{1}{2}\arctan(U/Q) \tag{5.9}$$

transforms to $\alpha - \phi$. Thus the orientation does not define a direction, it only refers the polarization to the (x, y)-plane.

The photon is peculiar in lacking a longitudinal polarization state, and the polarization is therefore not a vector in the (x, y)-plane; in fact it is a second-rank tensor. This is connected to the fact that the photon is massless. Recall that the theory of special relativity requires the photons to move with the speed of light in any frame. Therefore they must be massless, otherwise one would be able to accelerate them to higher speeds, or decelerate them to rest.

In a way, it appears as if there existed two kinds of photons. Physics has taken this into account by introducing an internal property, *spin*. Thus, one can talk about the two polarization states or about the two spin states of the photon. We shall come back to photon polarization later.

5.2 Adiabatic Expansion

For much of the thermal history of the Universe, the reaction rates of photons and other particles have been much greater than the Hubble expansion rate, so thermal equilibrium should have been maintained in any local comoving volume element $\mathrm{d}V$. There is then no net inflow or outflow of energy, which defines the expansion as adiabatic, as was done in Equation (4.25). The law of conservation of energy (4.24), also called the *first law of thermodynamics*, followed, by assuming that matter behaved as an expanding nonviscous fluid at constant pressure p.

Adiabaticity and Isentropy The entropy per unit comoving volume and physical volume $V = R^3$ at temperature T is defined by

$$S = \frac{1}{kT}(\rho c^2 + p)V. \tag{5.10}$$

Let us rewrite the first law of thermodynamics more generally in the form

$$\mathrm{d}[(\rho c^2 + p)V] = V\,\mathrm{d}p, \tag{5.11}$$

where the energy is $E = \rho c^2 V$.

The *second law of thermodynamics* can be written

$$\mathrm{d}S = \frac{1}{kT}[\mathrm{d}(\rho c^2 V) + p\,\mathrm{d}V]. \tag{5.12}$$

If the expansion is adiabatic and the pressure p is constant so that $\mathrm{d}(pV) = p\,\mathrm{d}V$, we recover Equation (4.25). Then it also follows that the expansion is *isentropic*:

$$\dot{S} = 0. \tag{5.13}$$

In the literature, the terms 'adiabaticity' and 'isentropy' are often confused.

Moreover, it follows from Equations (5.11) and the constancy of p that

$$\mathrm{d}E = -p\,\mathrm{d}V. \tag{5.14}$$

Thus a change in volume dV is compensated for by a change in energy dE at constant pressure and entropy.

The second law of thermodynamics states in particular that entropy cannot decrease in a closed system. The particles in a plasma possess maximum entropy when thermal equilibrium has been established. The assumption that the Universe expands adiabatically and isentropically is certainly very good during the radiation-dominated era when the fluid was composed of photons and elementary particles in thermal equilibrium.

This is also true during the matter-dominated era before matter clouds start to contract into galaxies under the influence of gravity. Even on a very large scale we may consider the galaxies forming a homogeneous 'fluid', an idealization as good as the cosmological principle that forms the basis of all our discussions. In fact, we have already relied on this assumption in the derivation of Einstein's equation and in the discussion of equations of state. However, the pressure in the 'fluid' of galaxies of density N is negligibly small, because it is caused by their random motion, just as the pressure in a gas is due to the random motion of the molecules. Since the average peculiar velocities $\langle v \rangle$ of the galaxies are of the order of $10^{-3}c$, the ratio of pressure $p = m\langle v \rangle^2 N$ to matter density ρ gives an equation of state (Problem 5) of the order of

$$w \approx \frac{m\langle v \rangle^2 N}{\rho c^2} = \frac{\langle v \rangle^2}{c^2} \approx 10^{-6}.$$

We have already relied on this value in the case of a matter-dominated universe when deriving Equation (4.30).

Radiation/Matter Domination. Let us compare the energy densities of radiation and matter. The energy density of electromagnetic radiation corresponding to one photon in a volume V is

$$\rho_{\mathrm{r}} c^2 \equiv \varepsilon_{\mathrm{r}} = \frac{h\nu}{V} = \frac{hc}{V\lambda}. \tag{5.15}$$

In an expanding universe with cosmic scale factor a, all distances scale as a and so does the wavelength λ. The volume V then scales as a^3; thus ε_{r} scales as a^{-4}. Here and in the following the subscript 'r' stands for radiation and relativistic particles, while 'm' stands for nonrelativistic (cold) matter.

Statistical mechanics tells us that the pressure in a nonviscous fluid is related to the energy density by the equation of state (4.32)

$$p = \tfrac{1}{3}\varepsilon, \tag{5.16}$$

where the factor $\frac{1}{3}$ comes from averaging over the three spatial directions. Thus pressure also scales as a^{-4}, so that it will become even more negligible in the future than it is now. The energy density of matter,

$$\rho_{\mathrm{m}} c^2 = \frac{mc^2}{V}, \tag{5.17}$$

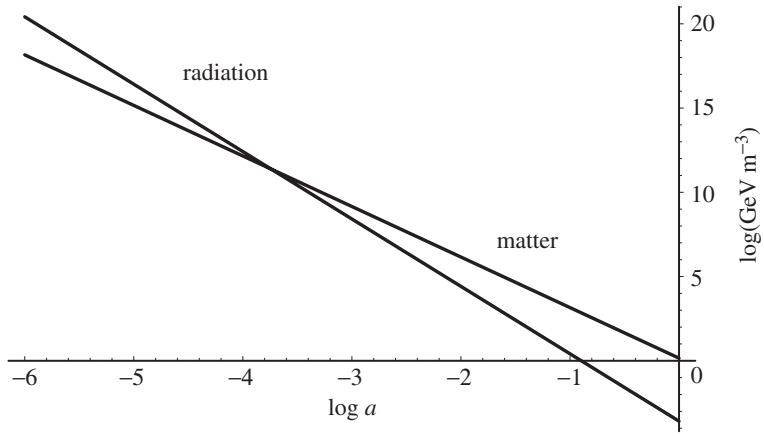


Figure 5.1 Scale dependence of the energy density in radiation ε_r, which dominates at small a, and in matter ρ_m, which dominates at large a, in units of $\log(\text{GeV m}^{-3})$. The scale value z_{eq} or a_{eq}^{-1} is evaluated in Equation (8.49) and indicated also in Figure 5.9.

also decreases with time, but only with the power a^{-3}. Thus the ratio of radiation energy to matter scales as a^{-1}:

$$\frac{\varepsilon_r}{\rho_m} \propto \frac{a^{-4}}{a^{-3}} \propto a^{-1}. \tag{5.18}$$

The present radiation energy density is predominantly in the form of microwaves and infrared light. Going backwards in time we reach an era when radiation and matter both contributed significantly to the total energy density. The change from radiation domination to matter domination is gradual: at $t = 1000$ yr the radiation fraction was about 90%, at $t = 2$ Myr only about 10% (see Figure 5.1). We shall calculate the time of equality t_{eq} in Chapter 8 when we know the present energy densities.

Temperature Dependence. A temperature T may be converted into units of energy by the dimensional relation

$$E = kT, \tag{5.19}$$

where k is the Boltzmann constant (Table A.2). Since E scales as a^{-1} it follows that also the temperature of radiation T_r scales as a^{-1} (Problem 1):

$$T_r \propto a^{-1} \propto (1 + z). \tag{5.20}$$

This dependence is roughly verified by measurements of the relic *cosmic microwave background* (CMB) radiation temperature at various times, corresponding to redshifts $z < 4.4$. Only two measurements give an absolute value: $T = 10 \pm 4$ K at $z = 2.338$ and $T = 12.1^{+1.70}_{-8.2}$ K at $z = 3.025$, in agreement with Equation (5.20).

Relativistic Particles. It is important to distinguish between relativistic and nonrelativistic particles because their energy spectra in thermal equilibrium are dif-

ferent. A coarse rule is that a particle is nonrelativistic when its kinetic energy is small in comparison with its mass, and relativistic when $E \gtrsim 10mc^2$. The masses of some cosmologically important particles are given in Table A.4. For comparison, the equivalent temperatures are also given. This gives a rough idea of the temperature of the heat bath when the respective particle is nonrelativistic.

The isentropy condition (5.13) can be applied to both relativistic and nonrelativistic particles. Let us first consider the relativistic particles which dominate the radiation era. Recall from Equation (2.69) that the energy of a particle depends on two terms, mass and kinetic energy,

$$E = \sqrt{m^2c^4 + P^2c^2}, \tag{5.21}$$

where P is momentum. For massless particles such as the photons, the mass term is of course absent; for relativistic particles it can be neglected.

Replacing E in Equation (5.14) by the energy density ε_r times the volume $a^3 \equiv V$, Equation (5.14) becomes

$$\mathrm{d}(a^3\varepsilon_\mathrm{r}) = -p\,\mathrm{d}(a^3). \tag{5.22}$$

Substituting ε_r for the pressure p from the equation of state (5.16) we obtain

$$a^3\,\mathrm{d}\varepsilon_\mathrm{r} + \varepsilon_\mathrm{r}\,\mathrm{d}a^3 = -\tfrac{1}{3}\varepsilon_\mathrm{r}\,\mathrm{d}a^3,$$

or

$$\frac{\mathrm{d}\varepsilon_\mathrm{r}}{\varepsilon_\mathrm{r}} = -\frac{4}{3}\frac{\mathrm{d}a^3}{a^3}. \tag{5.23}$$

The solution to this equation is

$$\varepsilon_\mathrm{r} \propto a^{-4}, \tag{5.24}$$

in agreement with our previous finding. We have in fact already used this result in Equation (4.33).

Non-relativistic Particles. For nonrelativistic particles the situation is different. Their kinetic energy ε_kin is small, so that the mass term in Equation (5.21) can no longer be neglected. The motion of n particles per unit volume is then characterized by a temperature T_m, causing a pressure

$$p = nkT_\mathrm{m}. \tag{5.25}$$

Note that T_m is not the temperature of matter in thermal equilibrium, but rather a bookkeeping device needed for dimensional reasons. The equation of state differs from that of radiation and relativistic matter, Equation (5.16), by a factor of 2:

$$p = \tfrac{2}{3}\varepsilon_\mathrm{kin}.$$

Including the mass term of the n particles, the energy density of nonrelativistic matter becomes

$$\rho_\mathrm{m} \equiv \varepsilon_\mathrm{m} = nmc^2 + \tfrac{3}{2}nkT_\mathrm{m}. \tag{5.26}$$

Substituting Equations (5.25) and (5.26) into Equation (5.22) we obtain

$$\mathrm{d}(a^3 nmc^2) + \tfrac{3}{2}\mathrm{d}(a^3 nkT_\mathrm{m}) = -nkT_\mathrm{m}\,\mathrm{d}a^3. \tag{5.27}$$

Let us assume that the total number of particles always remains the same: in a scattering reaction there are then always two particles coming in, and two going out, whatever their types. This is not strictly true because there also exist other types of reactions producing more than two particles in the final state. However, let us assume that the total number of particles in the volume V under consideration is $N = Vn$, and that N is constant during the adiabatic expansion,

$$\mathrm{d}N = \mathrm{d}(Vn) = \tfrac{4}{3}\pi\,\mathrm{d}(a^3 n) = 0. \tag{5.28}$$

The first term in Equation (5.27) then vanishes and we are left with

$$\tfrac{3}{2}a^3\mathrm{d}T_\mathrm{m} = -T_\mathrm{m}\,\mathrm{d}(a^3),$$

or

$$\frac{3}{2}\frac{\mathrm{d}T_\mathrm{m}}{T_\mathrm{m}} = -\frac{\mathrm{d}(a^3)}{a^3}.$$

The solution to this differential equation is of the form

$$T_\mathrm{m} \propto a^{-2}. \tag{5.29}$$

Thus we see that the temperature of nonrelativistic matter has a different dependence on the scale of expansion than does the temperature of radiation. This has profound implications for one of the most serious problems in thermodynamics in the 19th century.

Thermal Death. Suppose that the Universe starts out at some time with γ-rays at high energy and electrons at rest. This would be a highly ordered nonequilibrium system. The photons would obviously quickly distribute some of their energy to the electrons via various scattering interactions. Thus the original order would decrease, and the randomness or disorder would increase. The second law of thermodynamics states that any isolated system left by itself can only change towards greater disorder. The measure of disorder is entropy; thus the law says that entropy cannot decrease.

The counterexample which living organisms seem to furnish, since they build up ordered systems, is not valid. This is because no living organism exists in isolation; it consumes nutrients and produces waste. Thus, establishing that a living organism indeed increases entropy would require measurement of a much larger system, certainly not smaller than the Solar System.

It now seems to follow from the second law of thermodynamics that all energy would ultimately distribute itself evenly throughout the Universe, so that no further temperature differences would exist. The discoverer of the law of conservation of energy, *Hermann von Helmholtz* (1821–1894), came to the distressing

conclusion in 1854 that 'from this point on, the Universe will be falling into a state of eternal rest'. This state was named *thermal death*, and it preoccupied greatly both philosophers and scientists during the 19th century.

Now we see that this pessimistic conclusion was premature. Because, from the time when the temperatures of matter and radiation were equal,

$$T_{\rm m} = T_{\rm r},$$

we see from Equations (5.20) and (5.29) that the adiabatic expansion of the Universe causes matter to cool faster than radiation. Thus cold matter and hot radiation in an expanding Universe are not and will never be in thermal equilibrium on a cosmic timescale. This result permits us to solve the adiabatic equations of cold matter and hot radiation separately, as we in fact have.

5.3 Electroweak Interactions

Virtual Particles. In *quantum electrodynamics* (QED) the electromagnetic field is mediated by photons which are emitted by a charged particle and absorbed very shortly afterwards by another. Photons with such a brief existence during an interaction are called *virtual*, in contrast to real photons.

Virtual particles do not travel freely to or from the interaction region. Energy is not conserved in the production of virtual particles. This is possible because the energy imbalance arising at the creation of the virtual particle is compensated for when it is annihilated, so that the real particles emerging from the interaction region possess the same amount of energy as those entering the region. We have already met this argument in the discussion of Hawking radiation from black holes.

However, nature impedes the creation of very huge energy imbalances. For example, the masses of the *vector bosons* $\mathrm{W}^{\pm}$ and Z^0 mediating the electroweak interactions are almost 100 GeV. Reactions at much lower energies involving virtual vector bosons are therefore severely impeded, and much less frequent than electromagnetic interactions. For this reason such interactions are called *weak interactions*.

Real photons interact only with charged particles such as protons p, electrons e^- and their oppositely charged *antiparticles*, the *anti-proton* $\bar{\mathrm{p}}$ and the *positron* e^+. An example is the elastic *Compton scattering* of electrons by photons:

$$\gamma + \mathrm{e}^{\pm} \longrightarrow \gamma + \mathrm{e}^{\pm}. \tag{5.30}$$

As a result of virtual intermediate states neutral particles may exhibit electromagnetic properties such as magnetic moment.

Antimatter does not exist on Earth, and there is very little evidence for its presence elsewhere in the Galaxy. That does not mean that antiparticles are pure fiction: they are readily produced in particle accelerators and in violent astrophysical events. However, in an environment of matter, antiparticles rapidly meet their corresponding particles and annihilate each other. The asymmetry in the abundance

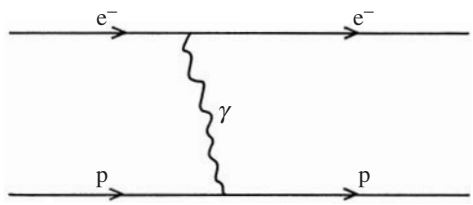


Figure 5.2 Feynman diagram for elastic scattering of an electron e^- on a proton p. This is an electromagnetic interaction mediated by a virtual photon γ. The direction of time is from left to right.

of matter and antimatter is surprising and needs an explanation. We shall deal with that in Section 6.7.

Charged particles interact via the electromagnetic field. Examples are the elastic scattering of electrons and positrons,

$$e^{\pm} + e^{\pm} \longrightarrow e^{\pm} + e^{\pm}, \tag{5.31}$$

and the Coulomb interaction between an electron and a proton, depicted in Figure 5.2. The free e^- and p enter the interaction region from the left, time running from left to right. They then exchange a virtual photon, and finally they leave the interaction region as free particles. This *Feynman diagram* does not show that the energies and momenta of the e^- and p change in the interaction. If one particle is fast and the other slow, the result of the interaction is that the slow particle picks up energy from the fast one, just as in the case of classical billiard balls. The Coulomb interaction between particles of like charges is repulsive and that between unlike charges is attractive. In both cases the energy and momentum get redistributed in the same way.

When an electron is captured by a free proton, they form a bound state, a hydrogen atom which is a very stable system. An electron and a positron may also form a bound atom-like state called *positronium*. This is a very unstable system: the electron and positron are antiparticles, so they rapidly end up annihilating each other according to the reaction

$$e^- + e^+ \longrightarrow \gamma + \gamma. \tag{5.32}$$

Since the total energy is conserved, the annihilation results in two (or three) photons possessing all the energy and flying away with it at the speed of light.

The reverse reaction is also possible. A photon may convert briefly into a virtual e^-e^+ pair, and another photon may collide with either one of these charged particles, knocking them out of the virtual state, thus creating a free electron–positron pair:

$$\gamma + \gamma \longrightarrow e^- + e^+. \tag{5.33}$$

This requires the energy of each photon to equal at least the electron (positron) mass, 0.51 MeV. If the photon energy is in excess of 0.51 MeV the e^-e^+ pair will not be created at rest, but both particles will acquire kinetic energy.

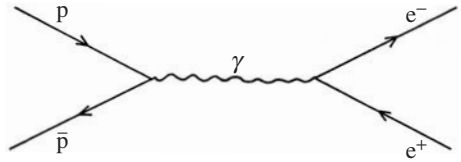


Figure 5.3 Feynman diagram for $p\bar{p}$ annihilation into e^+e^- via an intermediate virtual photon γ. The direction of time is from left to right.

Protons and anti-protons have electromagnetic interactions similar to positrons and electrons. They can also annihilate into photons, or for instance into an electron–positron pair via the mediation of a virtual photon,

$$p + \bar{p} \longrightarrow \gamma_{\text{virtual}} \longrightarrow e^- + e^+, \tag{5.34}$$

as depicted in Figure 5.3. The reverse reaction

$$e^- + e^+ \longrightarrow \gamma_{\text{virtual}} \longrightarrow p + \bar{p} \tag{5.35}$$

is also possible, provided the electron and positron possess enough kinetic energy to create a proton, or 938.3 MeV.

Note that the total electric charge is conserved throughout the reactions (5.30)–(5.35) and in Figures 5.1 and 5.2. Its value after the interaction (to the right of the arrow) is the same as it was before (to the left of the arrow). This is an important conservation law: electric charge can never disappear nor arise out of neutral vacuum. In the annihilation of an e^-e^+ pair into photons, all charges do indeed vanish, but only because the sum of the charges was zero to start with.

Baryons and Leptons. All the charged particles mentioned above have neutral partners as well. The partners of the p, $\bar{p}$, e^-, e^+ are the *neutron* n, the *anti-neutron* $\bar{n}$, the *electron neutrino* ν_e and the *electron anti-neutrino* $\bar{\nu}_e$. The p and n are called *nucleons*: they belong together with a host of excited nucleon-like states to the more general family of *baryons*. The $\bar{p}$ and $\bar{n}$ are correspondingly *anti-nucleons* or *anti-baryons*. The e^-, e^+, ν_e and $\bar{\nu}_e$ are called *leptons* and *anti-leptons* of the *electron family* (e).

We also have to introduce two more families or *flavours* of leptons: the μ family, comprising the charged *muons* $\mu^\pm$ and their associated neutrinos ν_μ, $\bar{\nu}_\mu$, and the τ family comprising $\tau^\pm$ and ν_τ, $\bar{\nu}_\tau$. The $\mu^\pm$ and $\tau^\pm$ are much more massive than the electrons, but otherwise their physics is very similar. They participate in reactions such as Equations (5.30)–(5.33) with e replaced by μ or τ, respectively. We shall discuss baryons and leptons in more detail in Chapter 6.

The charge can easily move from a charged particle to a neutral one as long as that does not violate the conservation of total charge in the reaction. Further, we need to know the following two conservation laws governing the behaviour of baryons and leptons.

(i) B or *baryon number* is conserved. This forbids the total number of baryons minus anti-baryons from changing in particle reactions. To help the book-

keeping in particle reactions we assign the value $B = 1$ to baryons and $B = -1$ to anti-baryons in a way analogous to the assignment of electric charges. Photons and leptons have $B = 0$.

(ii) L_l or *l-lepton number* is conserved for each of the flavours $l = \mathrm{e}, \mu, \tau$. This forbids the total number of l-leptons minus $\bar{l}$-anti-leptons from changing in particle reactions. We assign $L_\mathrm{e} = 1$ to e^- and ν_e, $L_\mathrm{e} = -1$ to e^+ and $\bar{\nu}_\mathrm{e}$, and correspondingly to the members of the μ and τ families. Photons and baryons have no lepton numbers.

However, there is an amendment to this rule, caused by the complications in the physics of neutrinos. Although the flavour state l is conserved in neutrino reactions, it is not conserved in free flight. To observe the flavour state l of neutrinos is not the same as observing the neutrino mass states. There are three neutrino mass states called ν_1, ν_2, ν_3, which are not identical to the flavour states; rather, they are quantum-mechanical superpositions of them. The states can mix in such a way that a pure mass state is a mixture of flavour states, and vice versa. Roughly, the ν_μ is the mixture of $\frac{1}{4}\nu_1$, $\frac{1}{4}\nu_2$ and $\frac{1}{2}\nu_3$.

All leptons participate in the weak interactions mediated by the heavy virtual vector bosons $\mathrm{W}^\pm$ and Z^0. The Z^0 is just like a photon except that it is very massive, about 91 GeV, and the $\mathrm{W}^\pm$ are its 10 GeV lighter charged partners. Weak leptonic reactions are

$$\mathrm{e}^\pm + \overset{(-)}{\nu}_\mathrm{e} \longrightarrow \mathrm{e}^\pm + \overset{(-)}{\nu}_\mathrm{e}, \tag{5.36}$$

$$\overset{(-)}{\nu}_\mathrm{e} + \overset{(-)}{\nu}_\mathrm{e} \longrightarrow \overset{(-)}{\nu}_\mathrm{e} + \overset{(-)}{\nu}_\mathrm{e}, \tag{5.37}$$

where $\overset{(-)}{\nu}_\mathrm{e}$ stands for ν_e or $\bar{\nu}_\mathrm{e}$. The Feynman diagrams of some of these reactions are shown in Figure 5.4. There is also the annihilation reaction

$$\mathrm{e}^- + \mathrm{e}^+ \longrightarrow \nu_\mathrm{e} + \bar{\nu}_\mathrm{e}, \tag{5.38}$$

and the pair production reaction

$$\nu_\mathrm{e} + \bar{\nu}_\mathrm{e} \longrightarrow \mathrm{e}^- + \mathrm{e}^+. \tag{5.39}$$

Similar reactions apply to the two other lepton families, replacing e above by μ or τ, respectively. Figure 5.4 illustrates that the total baryon number B and the total lepton number L_e are both conserved throughout the above reactions. Note that the ν_e can scatter against electrons by the two Feynman diagrams corresponding to $\mathrm{W}^\pm$ exchange and Z^0 exchange, respectively. In contrast, ν_μ and ν_τ can only scatter by the Z^0 exchange diagram, because of the separate conservation of lepton-family numbers.

Fermions and Bosons. The leptons and nucleons all have two spin states each. In the following we shall refer to them as *fermions*, after *Enrico Fermi* (1901–1954), whereas the photon and the W and Z are *bosons*, after *Satyendranath Bose*

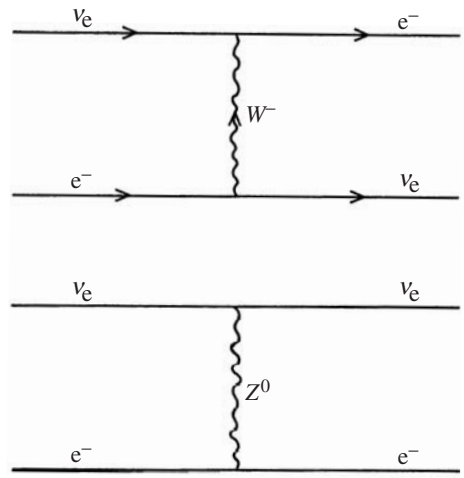


Figure 5.4 Feynman diagram for elastic scattering of an electron neutrino ν_e against an electron e^-. This weak interaction is mediated by a virtual W^- vector boson in the charged current reaction (upper figure), and by a virtual Z^0 vector boson in the neutral current reaction (lower figure). The direction of time is from left to right.

(1894–1974). In Table A.4 we have already met one more boson, the π meson, or *pion*. The difference between bosons and fermions is deep and fundamental. The number of spin states is even for fermions, odd for bosons (except the photon). They behave differently in a statistical ensemble. Fermions have antiparticles which most bosons do not. The *fermion number* is conserved, indeed separately for leptons and baryons, as we have seen. The number of bosons is not conserved; for instance, in pp collisions one can produce any number of pions and photons.

Two identical fermions refuse to get close to one another. This is the *Pauli exclusion force* responsible for the electron *degeneracy pressure* in white dwarfs and the neutron degeneracy pressure in neutron stars. A gas of free electrons will exhibit pressure even at a temperature of absolute zero. According to quantum mechanics, particles never have exactly zero velocity: they always carry out random motions, causing pressure. For electrons in a high-density medium such as a white dwarf with density $10^6\rho_\odot$, the degeneracy pressure is much larger than the thermal pressure, and it is enough to balance the pressure of gravity.

Bosons do not feel such a force, nothing inhibits them getting close to each other. However, it is beyond the scope and needs of this book to explain these properties further. They belong to the domains of quantum mechanics and quantum statistics.

The massive vector bosons $W^\pm$ and Z^0 have three spin or polarization states: the *transversal* (vertical and horizontal) states which the photons also have, and the *longitudinal state* along the direction of motion, which the photon is lacking.

In Table A.5 the number of spin states, $n_{\rm spin}$, of some of the cosmologically important particles are given. The fourth column tabulates $n_{\rm anti}$, which equals 2 for particles which possess a distinct antiparticle, otherwise it is equal to 1.

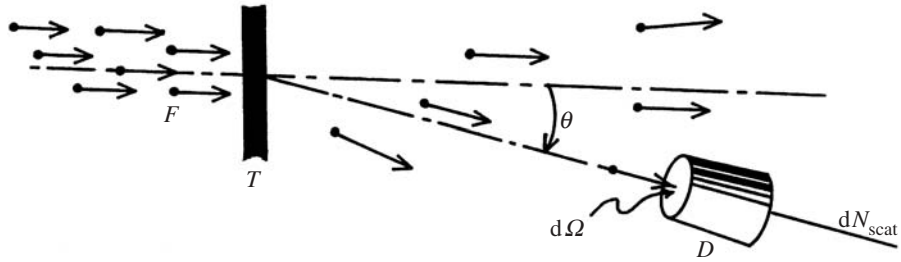


Figure 5.5 A beam of particles of flux F hitting a target T which scatters some of them into a detector D in the direction θ. The detector, which has a sensitive surface $d\Omega$, then records dN_{scat} scattered particles

As already explained, the number of distinct states or *degrees of freedom*, g, of photons in a statistical ensemble (in a plasma, say) is two. In general, due to the intricacies of quantum statistics, the degrees of freedom are the product of n_{spin}, n_{anti}, and a factor $n_{\text{Pauli}} = \frac{7}{8}$, which only enters for fermions obeying Fermi–Dirac statistics. For bosons this factor is unity. This product,

$$g = n_{\text{spin}} n_{\text{anti}} n_{\text{Pauli}}, \tag{5.40}$$

is tabulated in the fifth column of Table A.5.

Reaction Cross-Sections. The *rate* at which a reaction occurs, or the number of events per unit time, depends on the strength of the interaction as expressed by the coupling constant. It may also depend on many other details of the reaction, such as the spins and masses of the participating particles, and the energy E. All this information is contained in the reaction *cross-section*, σ. Let us follow an elementary argument to derive this quantity.

Suppose a beam contains k monoenergetic particles per m^3, all flying with velocity v m s^{-1} in the same direction (see Figure 5.5). This defines the *flux* F of particles per m^2 s in the beam. Let the beam hit a surface containing N target particles on which each beam particle may scatter. The number of particle reactions (actual scatterings) per second is then proportional to F and N. Consider the number of particles dN_{scat} scattered into a detector of angular opening $d\Omega$ in a direction θ from the beam direction (we assume azimuthal symmetry around the beam direction). Obviously, dN_{scat} is proportional to the number of particle reactions and to the detector opening,

$$dN_{\text{scat}} = FN\sigma(\theta)\, d\Omega.$$

The proportionality factor $\sigma(\theta)$ contains all the detailed information about the interaction. Integrating over all directions we can write this as

$$N_{\text{scat}} = FN\sigma, \tag{5.41}$$

where the proportionality constant

$$\sigma \equiv \int \sigma(\theta)\, d\Omega$$

has the dimension of a surface, here m^2. For this reason it has been named the *cross-section*.

One can also understand the reason for the surface units from a classical argument. Suppose a particle reaction can be treated as a game of billiard balls. Then the probability of hit is clearly proportional to the size of the target ball (of radius R), as seen by the hitting ball, or πR^2. The difference between billiard balls and particles is that σ should not be understood as the actual size, because that is not a useful quantity in quantum mechanics. Rather it depends on the interaction in a complicated manner.

The number density of relativistic particles other than photons is given by distributions very similar to the Planck distribution. Let us replace the photon energy $h\nu$ in Equation (5.3) by E, which is given by the relativistic expression (5.21). Noting that the kinematic variable is now the three-momentum $p = |\boldsymbol{p}|$ (since for relativistic particles we can ignore the mass), we can replace Planck's distribution by the number density of particle species i with momentum between p and $p + \mathrm{d}p$,

$$n_i(p)\,\mathrm{d}p = \frac{8\pi}{h^3}\frac{n_{\mathrm{spin},i}}{2}\frac{p^2\mathrm{d}p}{\mathrm{e}^{E_i(p)/kT_i} \pm 1}. \tag{5.42}$$

The $\pm$ sign is '$-$' for bosons and '$+$' for fermions, and the name for these distributions are the *Bose distribution* and the *Fermi distribution*, respectively. The Fermi distribution in the above form is actually a special case: it holds when the number of charged fermions equals the number of corresponding neutral fermions (the 'chemical potentials' vanish). In the following we shall need only that case.

The number density N of nonrelativistic particles of mass m is given by the *Maxwell–Boltzmann* distribution for an ideal, nondegenerate gas. Starting from Equation (5.42) we note that for nonrelativistic particles the energy kT is smaller than the mass, so that the term ± 1 in can be neglected in comparison with the exponential. Rewriting the Fermi distribution as a function of temperature rather than of momentum we obtain the Maxwell–Boltzmann distribution

$$N = n_{\mathrm{spin}}\frac{(2\pi mkT)^{3/2}}{(hc)^3}\,\mathrm{e}^{-E_i/kT_i}. \tag{5.43}$$

Note that because of the exponential term the number density falls exponentially as temperature falls. *James Clerk Maxwell* (1831–1879) was a contemporary of Stefan and Boltzmann.

5.4 The Early Radiation Era

Primordial Hot Plasma. In Section 5.1 we established the dependence of the number density of photons on temperature, N_γ in Equation (5.5), and the corresponding energy density, ε_{r} in Equation (5.6). For each species of relativistic fermions participating in the thermal equilibrium there is a specific number density. To find the total number density of particles sharing the available energy we have to count each particle species i weighted by the corresponding degrees of

freedom g_i. Remembering that $g_\gamma = 2$ for photons, we thus rewrite Equation (5.6) with a factor g_i explicitly visible:

$$\varepsilon_i = \tfrac{1}{2} g_i a_S T^4. \tag{5.44}$$

Remember that a_S is Stefan's constant. It turns out that this expression gives the correct energy density for every particle species if we insert its respective value of g_i from Table A.5.

Equation (5.5) can be correspondingly generalized to relativistic fermions. Their number density is

$$N_f = \tfrac{3}{4} N_\gamma. \tag{5.45}$$

In general, the primordial plasma was a mixture of particles, of which some are relativistic and some nonrelativistic at a given temperature. Since the number density of a nonrelativistic particle (given by the Maxwell–Boltzmann distribution, Equation (5.43)) is exponentially smaller than that of a relativistic particle, it is a good approximation to ignore nonrelativistic particles. Different species i with mass m_i have a number density which depends on m_i/T, and they may have a thermal distribution with a temperature T_i different from that of the photons. Let us define the *effective degrees of freedom* of the mixture as

$$g_* = \sum_{\text{bosons } i} g_i + \sum_{\text{fermions } j} g_j \left(\frac{T_j}{T}\right)^4. \tag{5.46}$$

As explained in the context of Equation (5.40) the sum over fermions includes a factor $\frac{7}{8}$, accounting for the difference between Fermi and Bose statistics. The factor $(T_j/T)^4$ applies only to neutrinos, which obtain a different temperature from the photons when they freeze out from the plasma (as we shall see later). Thus the energy density of the radiation in the plasma is

$$\varepsilon_r = \tfrac{1}{2} g_* a_S T^4. \tag{5.47}$$

Let us now derive a relation between the temperature scale and the timescale. We have already found the relation (4.40) between the size scale R and the timescale t during the radiation era,

$$a(t) \propto \sqrt{t}, \tag{5.48}$$

where we choose to omit the proportionality factor. The Hubble parameter can then be written

$$H = \frac{\dot{a}}{a} = \frac{1}{2t}. \tag{5.49}$$

Note that the proportionality factor omitted in Equation (5.48) has dropped out.

In Equation (4.35) we noted that the curvature term kc^2/R^2 in Friedmann's equations is negligibly small at early times during the radiation era. We then obtained the dynamical relation

$$\frac{\dot{a}}{a} = \left(\frac{8\pi G}{3}\rho\right)^{1/2}. \tag{5.50}$$

Inserting Equation (5.49) on the left and replacing the energy density ρ on the right by $\varepsilon_{\rm r}/c^2$, we find the relation sought between photon temperature and time:

$$\frac{1}{t} = \sqrt{\frac{16\pi G a_{\rm S}}{3c^2} g_*}\, T^2 = 3.07 \times 10^{-21} \sqrt{g_*} \frac{T^2}{[{\rm K}^2]} [{\rm s}^{-1}]. \tag{5.51}$$

The sum of degrees of freedom of a system of particles is of course the number of particles multiplied by the degrees of freedom per particle. Independently of the law of conservation of energy, the conservation of entropy implies that the energy is distributed equally between all degrees of freedom present in such a way that a change in degrees of freedom is accompanied by a change in random motion, or equivalently in temperature.

Thus entropy is related to order: the more degrees of freedom there are present, the more randomness or disorder the system possesses. When an assembly of particles (such as the molecules in a gas) does not possess energy other than kinetic energy (heat), its entropy is maximal when thermal equilibrium is reached. For a system of gravitating bodies, entropy increases by clumping, maximal entropy corresponding to a black hole.

Cooling Plasma. Let us now study the thermal history of the Universe during the radiation era. We may start at a time when the temperature was 10^{11} K, which corresponds to a mean energy of about 300 MeV. All of the electrons and photons then have an energy below the threshold for proton–anti-proton production (see Equation (5.35)). Thus the number of protons (and also neutrons and antinucleons) will no longer increase as a result of thermal collisions. They can only decrease, for a number of reasons which I shall explain.

Most of the other particles introduced in Section 5.3, the γ, e^-, μ^-, π, $\nu_{\rm e}$, ν_μ and ν_τ, as well as their antiparticles, are then present. The sum in Equation (5.46) is then, using the degrees of freedom in Table A.5,

$$g_* = 2 + 3 + 2 \times \tfrac{7}{2} + 3 \times \tfrac{7}{4} = \tfrac{69}{4}, \tag{5.52}$$

where the first term corresponds to the photon, the second to the three pions, the third to the two charged leptons and the fourth to the three kinds of neutrinos. Most of the unstable τ leptons disappeared shortly after 1.78 GeV. Inserting the value of g_* into Equation (5.51), we find that $t = 6.87$ ms at 10^{11} K.

We can follow the evolution of the function $g_*(T)$ in Figure 5.6. A comparison of the graph with Equation (5.52) shows that the latter is an underestimate. This is because all the particles in thermal equilibrium contribute, not only those accounted for in Equation (5.52), but also heavier particles which are thermalized by energetic photons and pions in the tail of their Boltzmann distributions. The steep drop at 200 MeV is caused by a phase transition: below 200 MeV we have hadronic matter (and leptons), whereas, above 200 MeV, the hadrons dissolve into their subconstituents, which contribute much more to g_*. We shall return to this in Section 6.6.

At this time the number density of nucleons decreases quickly because they have become nonrelativistic. Consequently, they have a larger probability of anni-

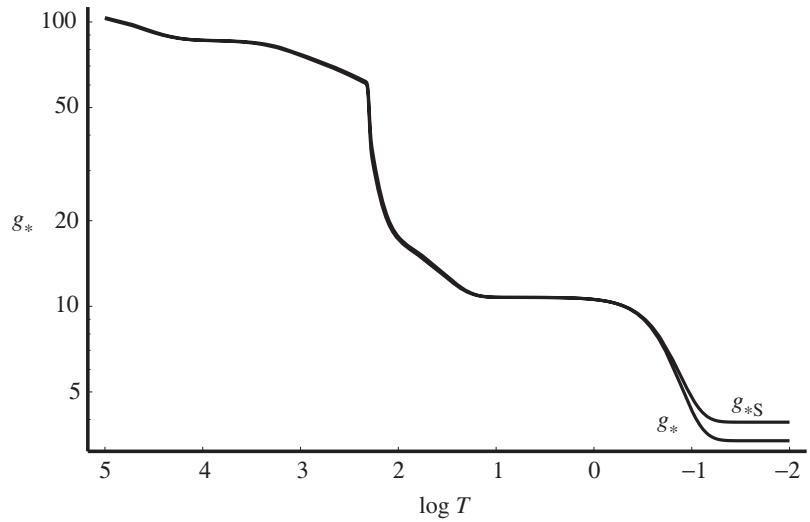


Figure 5.6 The evolution of the effective degrees of freedom contributing to the energy density, $g_*(T)$ and to the entropy density, $g_{*S}(T)$, as functions of $\log T$, where the temperature is in units of MeV [4].

hilating into lepton pairs, pion pairs or photons. Their number density is then no longer given by the Fermi distribution (5.42), but by the Maxwell–Boltzmann distribution, Equation (5.43). As can be seen from the latter, when T drops below the mass, the number density decreases rapidly because of the exponential factor. If there had been exactly the same number of nucleons and anti-nucleons, we would not expect many nucleons to remain to form matter. But, since we live in a matter-dominated Universe, there must have been some excess of nucleons early on. Note that neutrons and protons exist in equal numbers at the time under consideration.

Although the nucleons are very few, they still participate in electromagnetic reactions such as the elastic scattering of electrons,

$$\mathrm{e}^{\pm} + \mathrm{p} \longrightarrow \mathrm{e}^{\pm} + \mathrm{p}, \tag{5.53}$$

and in weak *charged current* reactions in which charged leptons and nucleons change into their neutral partners, and vice versa, as in

$$\mathrm{e}^{-} + \mathrm{p} \longrightarrow \nu_{\mathrm{e}} + \mathrm{n}, \tag{5.54}$$

$$\bar{\nu}_{\mathrm{e}} + \mathrm{p} \longrightarrow \mathrm{e}^{+} + \mathrm{n}. \tag{5.55}$$

Other such reactions are obtained by reversing the arrows, and by replacing $\mathrm{e}^{\pm}$ by $\mu^{\pm}$ or ν_{e} by ν_{μ} or ν_{τ}. The nucleons still participate in thermal equilibrium, but they are too few to play any role in the thermal history any more. This is why we could neglect them in Equation (5.52).

Below the pion mass (actually at about 70 MeV) the temperature in the Universe cools below the threshold for pion production:

$$(\mathrm{e}^{-} + \mathrm{e}^{+}) \text{ or } (\mu^{-} + \mu^{+}) \longrightarrow \gamma_{\text{virtual}} \longrightarrow \pi^{+} + \pi^{-}. \tag{5.56}$$

The reversed reactions, pion annihilation, still operate, reducing the number of pions. However, they disappear even faster by decay. This is always the fate when such lighter states are available, energy and momentum, as well as quantum numbers such as electric charge, baryon number and lepton numbers, being conserved. The pion, the muon and the tau lepton are examples of this. The pion decays mainly by the reactions

$$\pi^- \longrightarrow \mu^- + \bar{\nu}_\mu, \qquad \pi^+ \longrightarrow \mu^+ + \nu_\mu. \tag{5.57}$$

Thus g_* decreases by 3 to $\frac{57}{4}$. The difference in mass between the initial pion and the final state particles is

$$m_\pi - m_\mu - m_\nu = (139.6 - 105.7 - 0.0)\ \text{MeV} = 33.9\ \text{MeV}, \tag{5.58}$$

so 33.9 MeV is available as kinetic energy to the muon and the neutrino. This makes it very easy for the $\pi^\pm$ to decay, and in consequence its mean life is short, only 0.026 μs (the π^0 decays even faster). This is much less than the age of the Universe at 140 MeV, which is 23 μs from Equation (5.51). Note that the appearance of a charged lepton in the final state forces the simultaneous appearance of its anti-neutrino in order to conserve lepton number.

Also, the muons decay fast compared with the age of the Universe, with a lifetime of 2.2 μs, by the processes

$$\mu^- \longrightarrow \mathrm{e}^- + \bar{\nu}_\mathrm{e} + \nu_\mu, \qquad \mu^+ \longrightarrow \mathrm{e}^+ + \nu_\mathrm{e} + \bar{\nu}_\mu. \tag{5.59}$$

Almost the entire mass of the muon, or 105.7 MeV, is available as kinetic energy to the final state particles. This is the reason for its short mean life. Here again the conservation of lepton numbers, separately for the e-family and the μ-family, is observed.

Below the muon mass (actually at about 50 MeV), the temperature in the Universe cools below the threshold for muon-pair production:

$$\mathrm{e}^- + \mathrm{e}^+ \longrightarrow \gamma_\text{virtual} \longrightarrow \mu^+ + \mu^-. \tag{5.60}$$

The time elapsed is less than a millisecond. When the muons have disappeared, we can reduce g_* by $\frac{7}{2}$ to $\frac{43}{4}$.

From the reactions (5.57) and (5.59) we see that the end products of pion and muon decay are stable electrons and neutrinos. The lightest neutrino ν_1 is certainly stable, and the same is probably also true for ν_2 and ν_3. When this has taken place we are left with those neutrinos and electrons that only participate in weak reactions, with photons and with a very small number of nucleons. The number density of each lepton species is about the same as that of photons.

5.5 Photon and Lepton Decoupling

The considerations about which particles participate in thermal equilibrium at a given time depend on two timescales: the *reaction rate* of the particle, taking into account the reactions which are possible at that energy, and the *expansion rate* of the Universe. If the reaction rate is slow compared with the expansion rate, the distance between particles grows so fast that they cannot find each other.

Reaction Rates. The expansion rate is given by $H = \dot{a}/a$, and its temperature dependence by Equations (5.50) and (5.51). The average reaction rate can be written

$$\Gamma = \langle N v \sigma(E) \rangle, \tag{5.61}$$

where $\sigma(E)$ is the reaction cross-section (in units of m^2, say) as defined in Equation (5.41). The product of $\sigma(E)$ and the velocity v of the particle varies over the thermal distribution, so one has to average over it, as is indicated by the angle brackets. Multiplying this product by the number density N of particles per m^3, one obtains the mean rate Γ of reacting particles per second, or the mean collision time between collisions, Γ^{-1}.

The weak interaction cross-section turns out to be proportional to T^2,

$$\sigma \simeq \frac{G_F^2 (kT)^2}{\pi (\hbar c)^4}, \tag{5.62}$$

where G_F is the *Fermi coupling* measuring the strength of the weak interaction. The number density of the neutrinos is proportional to T^3 according to Equations (5.5) and (5.45). The reaction rate of neutrinos of all flavours then falls with decreasing temperature as T^5.

The condition for a given species of particle to remain in thermal equilibrium is then that the reaction rate Γ is larger than the expansion rate H, or equivalently that Γ^{-1} does not exceed the Hubble distance H^{-1},

$$\frac{\Gamma}{H} \gtrsim 1. \tag{5.63}$$

Inserting the T^5 dependence of the weak interaction rate Γ_{wi} and the T^2 dependence of the expansion rate H from Equation (5.51), we obtain

$$\frac{\Gamma_{wi}}{H} \propto T^3. \tag{5.64}$$

Thus there may be a temperature small enough that the condition (5.63) is no longer fulfilled.

Photon Reheating. Photons with energies below the electron mass can no longer produce e^+–e^- pairs, but the energy exchange between photons and electrons still continues by Compton scattering, reaction (5.30), or *Thomson scattering*, as it is called at very low energies. Electromagnetic cross-sections (subscript 'em') are proportional to T^{-2}, and the reaction rate is then proportional to T, so

$$\frac{\Gamma_{em}}{H} \propto \frac{1}{T}.$$

Contrary to the weak interaction case in Equation (5.64), the condition (5.63) is then satisfied for all temperatures, so electromagnetic interactions never freeze out. Electrons only decouple when they form neutral atoms during the *Recombination Era* and cease to scatter photons. The term recombination is slightly misleading, because the electrons have never been combined into atoms before. The

term comes from laboratory physics, where free electrons and *ionized* atoms are created by heating matter (and upon subsequent cooling the electrons and ions recombine into atoms) or from so-called HII regions, where interstellar plasma is ionized by ultraviolet radiation and characteristic *recombination radiation* is emitted when electrons and ions re-form.

The exothermic electron–positron annihilation, reaction (5.32), is now of mounting importance, creating new photons with energy 0.51 MeV. This is higher than the ambient photon temperature at that time, so the photon population gets reheated. To see just how important this reheating is, let us turn to the law of conservation of entropy.

Making use of the equation of state for relativistic particles (5.16), the entropy (5.10) can be written

$$S = \frac{4V}{3kT}\varepsilon_{\text{plasma}}.$$

Substituting the expression for $\varepsilon_{\text{plasma}}$ from Equation (5.47) one obtains

$$S = \frac{2g_*}{3}\frac{Va_S T^4}{kT}, \tag{5.65}$$

which is valid where we can ignore nonrelativistic particles. Now $a_S T^4$ is the energy density, so $Va_S T^4$ is energy, just like kT, and thus $Va_S T^4/kT$ is a constant. g_* is also a constant, except at the thresholds where particle species decouple. The physical meaning of entropy of a system is really its degrees of freedom multiplied by some constant, as one sees here. In Equation (5.5) we saw that the entropy density can also be written

$$s \equiv \frac{S}{V} = \tfrac{3}{2}\zeta(3)g_\gamma N_\gamma, \tag{5.66}$$

where N_γ is the number density of photons. Between two decoupling thresholds we then have

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{2g_*}{3}\frac{Va_S T^3}{k}\right) = 0. \tag{5.67}$$

The second law of thermodynamics requires that entropy should be conserved in reversible processes, also at thresholds where g_* changes. This is only possible if T also changes in such a way that $g_* T^3$ remains a constant. When a relativistic particle becomes nonrelativistic and disappears, its entropy is shared between the particles remaining in thermal contact, causing some slight slowdown in the cooling rate. Photons never become nonrelativistic; neither do the practically massless neutrinos, and therefore they continue to share the entropy of the Universe, each species conserving its entropy separately.

Let us now apply this argument to the situation when the positrons and most of the electrons disappear by annihilation below 0.2 MeV. We denote temperatures and entropies just above this energy by a subscript '+', and below it by '−'. Above this energy, the particles in thermal equilibrium are γ, e^-, e^+. Then the entropy

$$S = \frac{2}{3}\left(2 + \frac{7}{2}\right)\frac{Va_S T_+^3}{k}. \tag{5.68}$$

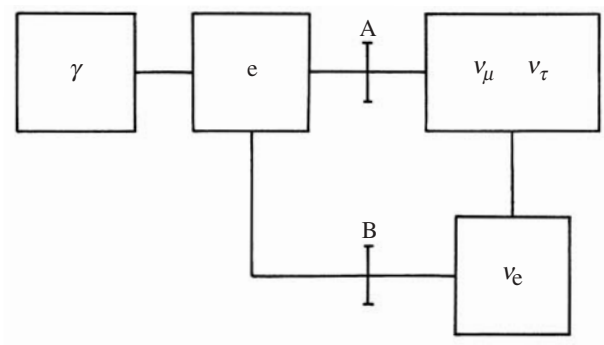


Figure 5.7 A system of communicating vessels illustrating particles in thermal equilibrium (from K. Kainulainen, unpublished research). At 3.7 MeV, valve A closes so that ν_μ and ν_τ decouple. At 2.3 MeV, valve B closes so that ν_e also decouples, leaving only e^- and γ in thermal contact.

Below that energy, only photons contribute the factor $g_* = 2$. Consequently, the ratio of entropies S_+ and S_- is

$$\frac{S_+}{S_-} = \frac{11}{4}\left(\frac{T_+}{T_-}\right)^3. \tag{5.69}$$

But entropy must be conserved so this ratio must be unity. It then follows that

$$T_- = \left(\tfrac{11}{4}\right)^{1/3} T_+ = 1.40 T_+. \tag{5.70}$$

Thus the temperature T_γ of the photons increases by a factor 1.40 as the Universe cools below the threshold for electron–positron pair production. Actually, the temperature increase is so small and so gradual that it only slows down the cooling rate temporarily.

Neutrino Decoupling. When the neutrinos no longer obey the condition (5.63) they *decouple* or *freeze out* from all interactions, and begin a free expansion. The decoupling of ν_μ and ν_τ occurs at 3.5 MeV, whereas the ν_e decouple at 2.3 MeV. This can be depicted as a set of connecting baths containing different particles, and having valves which close at given temperatures (see Figure 5.7).

At decoupling, the neutrinos are still relativistic, since they are so light (Table A.3). Thus their energy distribution is given by the Fermi distribution, Equation (5.42), and their temperature equals that of the photons, $T_\nu = T_\gamma$, decreasing with the increasing scale of the Universe as a^{-1}. But the neutrinos do not participate in the reheating process, and they do not share the entropy of the photons, so from now on they remain colder than the photons:

$$T_\nu = T_\gamma / 1.40. \tag{5.71}$$

The number density N_ν of neutrinos can be calculated as in Equation (5.5) using Equation (5.3), except that the -1 term in Equation (5.3) has to be replaced by

+1, which is required for fermions (see Equation (5.42)). In the number density distributions (5.3) and (5.42), we have ignored possible chemical potentials for all fermions, which one can do for a thermal radiation background; for neutrinos it is an unproven assumption that nonetheless appears in reasonable agreement with their oscillation parameters.

The result is that N_ν is a factor of $\frac{3}{4}$ times N_γ at the same temperature. Taking the difference between temperatures T_ν and T_γ into account and noting from Equation (5.5) that N_γ is proportional to T^3, one finds

$$N_\nu = \tfrac{3}{4}\tfrac{4}{11} N_\gamma. \tag{5.72}$$

After decoupling, the neutrino contribution to g_* decreases because the ratio T_i/T in Equation (5.46) is now less than one. Thus the present value is

$$g_*(T_0) = 2 + 3\tfrac{7}{4}\left(\tfrac{4}{11}\right)^{4/3} = 3.36. \tag{5.73}$$

The entropy density also depends on g_*, but now the temperature dependence for the neutrino contribution in Equation (5.46) is $(T_i/T)^3$ rather than a power of four. The effective degrees of freedom are in that case given by Equation (5.73) if the power $\frac{4}{3}$ is replaced by 1. This curve is denoted g_{*S} in Figure 5.6.

The density parameter is

$$\Omega_\nu = \frac{3}{11}\frac{N_\gamma}{\rho_c h^2}\sum_i m_i. \tag{5.74}$$

Recombination Era. As long as there are free electrons, the primordial photons are thermalized by Thomson scattering against them, and this prohibits the electrons from decoupling, in contrast to neutrinos. Each scattering polarizes the photons, but on average this is washed out. The electromagnetic reaction rate is much higher than the weak reaction rate of the neutrinos; in fact, it is higher than the expansion rate of the Universe, so the condition (5.63) is fulfilled.

Eventually the Universe expands and cools to such an extent, to about 1000 K, that electrons are captured into atomic orbits, primarily by protons but also by the trace amounts of ionized helium and other light nuclei. This process is referred to as recombination. Unlike the unstable particles n, π, μ that decay spontaneously liberating kinetic energy in exothermic reactions, the hydrogen atom H is a *bound state* of a proton and an electron. Its mass is less than the p and e^- masses together,

$$m_H - m_p - m_e = -13.59 \text{ eV}, \tag{5.75}$$

so it cannot disintegrate spontaneously into a free proton and a free electron. The mass difference (5.75) is the *binding energy* of the hydrogen atom.

The physics of recombination is somewhat subtle. Initially one might think that recombination occurs when the photon temperature drops below 13.59 eV, making formation of neutral hydrogen energetically favourable. Two characteristics of the physics push the recombination temperature lower. The first, and the easiest

to elaborate, is that there are vastly more photons than electrons and so in thermal equilibrium even a small proportion of high-energy photons are sufficient to maintain effectively complete ionization. Photons in thermal equilibrium have the blackbody spectrum given by equation (5.3). Even for photon temperatures somewhat below 13.59 eV there will be enough highly energetic photons in the Wein tail (as the high-energy section is termed) to ionize any neutral hydrogen. The large amount of entropy in the Universe also favours free protons and electrons.

With respect to the thermal history of the Universe, the fact that photons do not scatter against neutral atoms is critical. As recombination proceeds and the number of electrons falls, matter and radiation decouple. This has two results. First, with matter and radiation no longer in thermal equilibrium the thermal history of the two proceed independently. Perturbations in matter are no longer damped by interaction with radiation and any such perturbations can grow into structures through gravitational instability. Decoupling thus initiates the period of structure formation that has led to our present Universe being populated with stars, galaxies, galaxy clusters, etc.

The second result is that, with photons no longer scattering against a sea of electrons, the photons can stream freely through the Universe; upon recombination, the Universe becomes transparent to light. Prior to recombination, the Universe was opaque to electromagnetic radiation (although not to neutrinos) and it would have been impossible to do astronomy if this situation had persisted until today. The freely streaming photons from this era form the CMB radiation and their point of last contact with matter forms a spherical shell called the *last scattering surface* (LSS). The era of recombination provides a crucial observational limit beyond which we cannot hope to see using electromagnetic radiation.

The LSS is not a sharp boundary and does not exist at a unique redshift: it is actually a thin shell. The photons from the LSS preserve the polarization they incurred in the last Thomson scattering. This remaining primordial polarization is an interesting detectable signal, albeit much weaker than the intensity of the thermalized radiation (we shall return to discuss this further in Section 8.3).

The LSS of the Universe has an exact analogue in the surface of the Sun. Photons inside the Sun are continuously scattered, so it takes millions of years for some photons to reach the surface. But once they do not scatter any more they continue in straight lines (really on geodesics) towards us. Therefore, we can see the surface of the Sun, which is the LSS of the solar photons, but we cannot see the solar interior. We can also observe that sunlight is linearly polarized. In contrast, neutrinos hardly scatter at all in the Sun, thus neutrino radiation brings us a clear (albeit faint with present neutrino detectors) picture of the interior of the Sun.

Equilibrium Theory. An analysis based on thermal equilibrium, the *Saha equation*, implies that the temperature must fall to about 0.3 eV before the proportion of high-energy photons falls sufficiently to allow recombination to occur. The Saha analysis also implies that the time (or energy or redshift) for decoupling and last scattering depend on cosmological parameters such as the total cosmic density parameter Ω_0, the baryon density Ω_b, and the Hubble parameter. However, a sec-

ond feature of the physics of recombination implies that the equilibrium analysis itself is not sufficient.

The Saha analysis describes the initial phase of departure from complete ionization but, as recombination proceeds, the assumption of equilibrium ceases to be appropriate (see, for example, [1]). Paradoxically, the problem is that electromagnetic interactions are too fast (in contrast with the weak interaction that freezes out from equilibrium because of a small cross-section). A single recombination directly to the ground state would produce a photon with energy greater than the 13.59 eV binding energy and this photon would travel until it encountered a neutral atom and ionized it. This implies that recombination in an infinite static universe would have to proceed by smaller intermediate steps (thus not directly to the ground state).

In fact the situation is even worse, because reaching the ground state by single photon emission requires transition from the 2P to 1S levels and thus production of photons with energy at least 10.2 eV (Lyman α with $\lambda = 1216$ Å). As these photons become abundant they will re-ionize any neutral hydrogen through multiple absorption and so it would seem that recombination will be, at a minimum, severely impeded. (Recombination in a finite HII region is different because the Lyα photons can escape (see [1, p. 286]).)

There is an alternative path, however. Two-photon emission generated by the $2S \rightarrow 1S$ transition produces lower-energy photons. The process is slow (with a lifetime of approximately 0.1 s), so recombination proceeds at a rate quite different from the Saha prediction. Consequently, all the times predicted by this nonequilibrium analysis differs notably from the Saha prediction, but, interestingly, in such a way that the times of decoupling and last scattering have practically no dependence on cosmological parameters.

Summary. Recombination, decoupling and last scattering do not occur at the exactly same time. It should also be noted that these terms are often used interchangeably in the literature, so what we refer to as the LSS may also be called the time of recombination or decoupling. Approximate results for these times, following from Peacock [1], are summarized below.

Recombination is defined as the time when 90% of the electrons have combined into neutral atoms. This occurred at redshift

$$a_{\rm rec}^{-1} = 1 + z_{\rm rec} \approx 1310\left(\frac{\Omega_{\rm b}h}{\sqrt{\Omega_0}}\right)^{0.078} \approx 910\text{–}1340. \tag{5.76}$$

Last scattering is defined as the time when photons start to stream freely. This occurred at redshift

$$a_{\rm LSS}^{-1} = 1 + z_{\rm LSS} = 1065 \pm 80, \tag{5.77}$$

when the Universe was $180\,000(\Omega_0 h^2)^{-1/2}$ years old and at a temperature of 0.26 eV, thus right after the recombination time. This redshift has been determined much more precisely by WMAP [5], as we shall show in Section 8.4.

Decoupling is defined as the time when the reaction rate (scattering) falls below the expansion rate of the Universe and matter falls out of thermal equilibrium with photons. This occurred at redshift

$$a_{\rm dec}^{-1} = 1 + z_{\rm dec} \approx 890, \tag{5.78}$$

when the Universe was some 380 000 years old. All three events depend on the number of free electrons (the ionization fraction) but in slightly different ways. As a result these events do not occur at exactly the same time.

5.6 Big Bang Nucleosynthesis

Let us now turn to the fate of the remaining nucleons. Note that the charged current reactions (5.54) and (5.55) changing a proton to a neutron are *endothermic*: they require some input energy to provide for the mass difference. In reaction (5.54) this difference is 0.8 MeV and in reaction (5.55) it is 1.8 MeV (use the masses in Table A.4!). The reversed reactions are *exothermic*. They liberate energy and they can then always proceed without any threshold limitation.

The neutrons and protons are then nonrelativistic, so their number densities are each given by Maxwell–Boltzmann distributions (5.43). Their ratio in equilibrium is given by

$$\frac{N_{\rm n}}{N_{\rm p}} = \left(\frac{m_{\rm n}}{m_{\rm p}}\right)^{3/2} \exp\left(-\frac{m_{\rm n} - m_{\rm p}}{kT}\right). \tag{5.79}$$

At energies of the order of $m_{\rm n} - m_{\rm p} = 1.293$ MeV or less, this ratio is dominated by the exponential. Thus, at $kT = 0.8$ MeV, the ratio has dropped from 1 to $\frac{1}{5}$. As the Universe cools and the energy approaches 0.8 MeV, the endothermic neutron-producing reactions stop, one by one. Then no more neutrons are produced but some of those that already exist get converted into protons in the exothermic reactions.

Nuclear Fusion. Already, at a few MeV, nuclear *fusion reactions* start to build up light elements. These reactions are exothermic: when a neutron and a proton fuse into a bound state some of the nucleonic matter is converted into pure energy according to Einstein's formula (2.68). This binding energy of the *deuteron* d,

$$m_{\rm p} + m_{\rm n} - m_{\rm d} = 2.22 \text{ MeV}, \tag{5.80}$$

is liberated in the form of radiation:

$$\rm n + p \longrightarrow d + \gamma. \tag{5.81}$$

The deuteron is also written $^2\rm H^+$ in general nuclear physics notation, where the superscript $A = 2$ indicates the number of nucleons and the electric charge is given by the superscript '+'. The bound state formed by a deuteron and an electron is the *deuterium* atom ^{2}H, which of course is electrically neutral. Although the

deuterons are formed in very small quantities, they are of crucial importance to the final composition of matter.

As long as photons of 2.22 MeV or more are available, the reaction (5.81) can go the other way: the deuterons *photodisintegrate* into free protons and neutrons. Even when the mean temperature of radiation drops considerably below 2.22 MeV, there is still a high-energy tail of the Planck distribution containing hard γ-rays which destroy the deuterons as fast as they are produced.

All evidence suggests that the number density of baryons, or equivalently nucleons, is today very small. In particular, we are able to calculate it to within a factor $\Omega_{\rm B}h^2$,

$$N_{\rm B} = \frac{\rho_{\rm B}}{m_{\rm B}} = \frac{\Omega_{\rm B}\rho_{\rm c}}{m_{\rm B}} \simeq 11.3\Omega_{\rm B}h^2\ {\rm m}^{-3}. \tag{5.82}$$

At the end of this section we shall discuss the value of the baryon density parameter $\Omega_{\rm B}$, which is a few per cent.

The photon number density today is $N_\gamma = 4.11\times10^8$ per m^3 from Equation (5.5). It is clear then that $N_{\rm B}/N_\gamma$ is such a small figure that only an extremely tiny fraction of the high-energy tail of the photon distribution may contain sufficiently many hard γ-rays to photodisintegrate the deuterons. However, the 2.22 MeV photons created in photodisintegration do not thermalize, so they will continue to photodisintegrate deuterium until they have been redshifted below this threshold. Another obstacle to permanent deuteron production is the high entropy per nucleon in the Universe. Each time a deuteron is produced, the degrees of freedom decrease, and so the entropy must be shared among the remaining nucleons. This raises their temperature, counteracting the formation of deuterons. Detailed calculations show that deuteron production becomes thermodynamically favoured only at 0.07 MeV. Thus, although deuterons are favoured on energetic grounds already at 2 MeV, free nucleons continue to be favoured by the high entropy down to 0.07 MeV.

Other nuclear fusion reactions also commence at a few MeV. The npp bound state $^3{\rm He}^{++}$ is produced in the fusion of two deuterons,

$$\mathrm{d} + \mathrm{d} \longrightarrow {}^3\mathrm{He}^{++} + \mathrm{n}, \tag{5.83}$$

$$\mathrm{p} + \mathrm{d} \longrightarrow {}^3\mathrm{He}^{++} + \gamma, \tag{5.84}$$

where the final-state particles share the binding energy

$$2m_{\rm p} + m_{\rm n} - m({}^3{\rm He}^{++}) = 7.72\ {\rm MeV}. \tag{5.85}$$

This reaction is also hampered by the large entropy per nucleon, so it becomes thermodynamically favoured only at 0.11 MeV.

The nnp bound state $^3{\rm H}^+$, or *triton* t, is the ionized *tritium* atom, ^{3}H. It is produced in the fusion reactions

$$\mathrm{n} + \mathrm{d} \longrightarrow \mathrm{t} + \gamma, \tag{5.86}$$

$$\mathrm{d} + \mathrm{d} \longrightarrow \mathrm{t} + \mathrm{p}, \tag{5.87}$$

$$\mathrm{n} + {}^3\mathrm{He} \longrightarrow \mathrm{t} + \mathrm{p}, \tag{5.88}$$

with the binding energy

$$m_{\mathrm{p}} + 2m_{\mathrm{n}} - m_{\mathrm{t}} = 8.48 \text{ MeV}. \tag{5.89}$$

A very stable nucleus is the nnpp bound state $^4\mathrm{He}^{++}$ with a very large binding energy,

$$2m_{\mathrm{p}} + 2m_{\mathrm{n}} - m(^4\mathrm{He}^{++}) = 28.3 \text{ MeV}. \tag{5.90}$$

Once its production is favoured by the entropy law, at about 0.28 MeV, there are no more γ-rays left that are hard enough to photodisintegrate it. From the examples set by the deuteron fusion reactions above, it may seem that $^4\mathrm{He}^{++}$ would be most naturally produced in the reaction

$$\mathrm{d} + \mathrm{d} \longrightarrow {}^4\mathrm{He}^{++} + \gamma. \tag{5.91}$$

However, $^3\mathrm{He}^{++}$ and $^3\mathrm{H}^{+}$ production is preferred over deuteron fusion, so $^4\mathrm{He}^{++}$ is only produced in a second step when these nuclei become abundant. The reactions are then

$$\mathrm{n} + {}^3\mathrm{He}^{++} \longrightarrow {}^4\mathrm{He}^{++} + \gamma, \tag{5.92}$$

$$\mathrm{d} + {}^3\mathrm{He}^{++} \longrightarrow {}^4\mathrm{He}^{++} + \mathrm{p}, \tag{5.93}$$

$$\mathrm{p} + \mathrm{t} \longrightarrow {}^4\mathrm{He}^{++} + \gamma, \tag{5.94}$$

$$\mathrm{d} + \mathrm{t} \longrightarrow {}^4\mathrm{He}^{++} + \mathrm{n}. \tag{5.95}$$

The delay before these reactions start is often referred to as the *deuterium bottleneck*.

Below 0.8 MeV occasional weak interactions in the high-energy tails of the lepton and nucleon Fermi distributions reduce the n/p ratio further, but no longer by the exponential factor in Equation (5.79). The neutrons also decay into protons by *beta decay*,

$$\mathrm{n} \longrightarrow \mathrm{e}^- + \bar{\nu}_{\mathrm{e}} + \mathrm{p}, \tag{5.96}$$

liberating

$$m_{\mathrm{n}} - m_{\mathrm{p}} - m_{\mathrm{e}} - m_{\nu} = 0.8 \text{ MeV} \tag{5.97}$$

of kinetic energy in the process. This amount is very small compared with the neutron mass of 939.6 MeV. In consequence the decay is inhibited and very slow: the neutron mean life is 887 s. In comparison with the age of the Universe, which at this time is a few tens of seconds, the neutrons are essentially stable. The protons are stable even on scales of billions of years, so their number is not going to decrease by decay.

At 0.1 MeV, when the temperature is 1.2×10^9 K and the time elapsed since the Big Bang is a little over two minutes, the beta decays have reduced the neutron/proton ratio to its final value:

$$\frac{N_{\mathrm{n}}}{N_{\mathrm{p}}} \simeq \frac{1}{7}. \tag{5.98}$$

The temperature dependence of this ratio, as well as the equilibrium (Maxwell–Boltzmann) ratio, is shown in Figure 5.8.

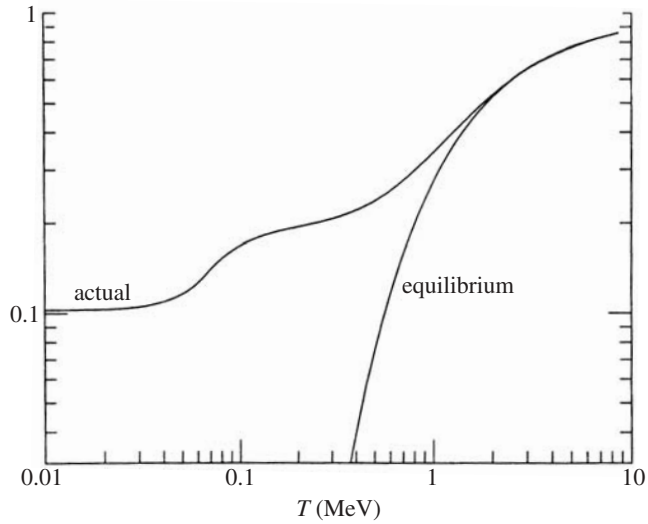


Figure 5.8 The equilibrium and actual values of the *n/p* ratio. Courtesy of E. W. Kolb and M. S. Turner.

These remaining neutrons have no time to decay before they fuse into deuterons and subsequently into $^4He^{++}$. There they stayed until today because bound neutrons do not decay. The same number of protons as neutrons go into ^4He, and the remaining free protons are the nuclei of future hydrogen atoms. Thus the end result of the nucleosynthesis taking place between 100 and 700 s after the Big Bang is a Universe composed almost entirely of hydrogen and helium ions. But why not heavier nuclei?

It is an odd circumstance of nature that, although there exist stable nuclei composed of $A = 1$, 2, 3 and 4 nucleons, no nucleus of $A = 5$, or of $A = 8$, exists. In between these gaps, there exist the unstable nuclei ^6Li and ^7Be, and the stable ^7Li. Because of these gaps and because ^4He is so strongly bound, nucleosynthesis essentially stops after ^4He production. Only minute quantities of the stable nuclei ^2H, ^3He and ^7Li can be produced.

The fusion rate at energy E of two nuclei of charges Z_1, Z_2 is proportional to the *Gamow penetration factor*

$$\exp\left(-\frac{2Z_1Z_2}{\sqrt{E}}\right). \tag{5.99}$$

Thus as the energy decreases, the fusion of nuclei other than the very lightest ones becomes rapidly improbable.

Relic ^{4}He Abundance. The relic abundances of the light elements bear an important testimony of the n/p ratio at the time of the nucleosynthesis when the Universe was only a few minutes old. In fact, this is the earliest testimony of the Big Bang we have. Recombination occurred some 300 000 years later, when the stable ions captured all the electrons to become neutral atoms. The CMB testimony is

from that time. There is also more recent information available in galaxy cluster observations from $z < 0.25$.

From the ratio (5.98) we obtain immediately the ratio of ^{4}He to ^{1}H:

$$X_4 \equiv \frac{N(^4\mathrm{He})}{N(^1\mathrm{H})} = \frac{N_\mathrm{n}/2}{N_\mathrm{p} - N_\mathrm{n}} \simeq \frac{1}{12}. \tag{5.100}$$

The number of ^{4}He nuclei is clearly half the number of neutrons when the minute amounts of ^{2}H, ^{3}He and ^{7}Li are neglected. The same number of protons as neutrons go into ^{4}He, thus the excess number of protons becoming hydrogen is $N_\mathrm{p} - N_\mathrm{n}$. The ratio of mass in ^{4}He to total mass in ^{1}H and ^{4}He is

$$Y_4 \equiv \frac{4X_4}{1 + 4X_4} \simeq 0.25. \tag{5.101}$$

This is a function of the ratio of baryons to photons

$$\eta \equiv \frac{N_\mathrm{b}}{N_\gamma} \simeq 2.75 \times 10^{-8} \Omega_\mathrm{b} h^2, \tag{5.102}$$

using N_γ from Table A.6.

The helium mass abundance Y_4 depends sensitively on several parameters. If the number of baryons increases, Ω_b and η also increase, and the entropy per baryon decreases. Since the large entropy per baryon was the main obstacle to early deuteron and helium production, the consequence is that helium production can start earlier. But then the neutrons would have had less time to β-decay, so the neutron/proton ratio would be larger than $\frac{1}{7}$. It follows that more helium will be produced: Y_4 increases.

The abundances of the light elements also depend on the neutron mean life τ_n and on the number of neutrino families F_ν, both of which were poorly known until 1990. Although τ_n is now known to 1‰ [2], and F_ν is known to be 3 ± 4‰ [2], it may be instructive to follow the arguments about how they affect the value of Y_4.

Let us rewrite the decoupling condition (5.64) for neutrons

$$\frac{\Gamma_\mathrm{wi}}{H} = A T_\mathrm{d}^3, \tag{5.103}$$

where A is the proportionality constant left out of Equation (5.64) and T_d is the decoupling temperature. An increase in the neutron mean life implies a decrease in the reaction rate Γ_wi and therefore a decrease in A. At temperature T_d the ratio of the reaction rate to the expansion rate is unity; thus

$$T_\mathrm{d} = A^{-1/3}. \tag{5.104}$$

Hence a longer neutron mean life implies a higher decoupling temperature and an earlier decoupling time. As we have already seen, an earlier start of helium production leads to an increase in Y_4.

The expansion rate H of the Universe is, according to Equations (5.49) and (5.51), proportional to $\sqrt{g_*}$, which in turn depends on the number of neutrino families F_ν. In Equations (5.52) we had set $F_\nu = 3$. Thus, if there were more than three neutrino families, H would increase and A would decrease with the same consequences as in the previous example. Similarly, if the number of neutrinos were very different from the number of anti-neutrinos, contrary to the assumptions in standard Big Bang cosmology, H would also increase.

Light Element Abundance Observations. The value of $\Omega_b h^2$ (or η) is obtained in direct measurements of the relic abundances of ^{4}He, ^{3}He, ^{2}H or D, and ^{7}Li from the time when the Universe was only a few minutes old. Although the ^{4}He mass ratio Y_4 is 0.25, the ^{3}He and ^{2}H mass ratios are less than 10^{-4} and the ^{7}Li mass ratio as small as a few times 10^{-10}, they all agree remarkably well on a common value for η.

If the observed abundances are indeed of cosmological origin, they must not be affected significantly by later stellar processes. The helium isotopes ^{3}He and ^{4}He cannot be destroyed easily but they are continuously produced in stellar interiors. Some recent helium is blown off from supernova progenitors, but that fraction can be corrected for by observing the total abundance in hydrogen clouds of different ages and extrapolating to time zero. The remainder is then primordial helium emanating from BBN. On the other hand, the deuterium abundance can only decrease; it is easily burned to ^{3}He in later stellar events. The case of ^{7}Li is complicated because some fraction is due to later galactic cosmic ray spallation products.

The ^{4}He abundance is easiest to observe, but it is also least sensitive to $\Omega_b h^2$, its dependence is logarithmic, so only very precise measurements are relevant. The best 'laboratories' for measuring the ^{4}He abundance are a class of low-luminosity dwarf galaxies called blue compact dwarf (BCD) galaxies, which undergo an intense burst of star formation in a very compact region. The BCDs are among the most metal-deficient gas-rich galaxies known (astronomers call all elements heavier than helium *metals*). Since their gas has not been processed through many generations of stars, it should approximate well the pristine primordial gas.

The ^{3}He isotope can be seen in galactic star-forming regions containing ionized hydrogen (HII), in the local interstellar medium and in planetary nebulae. Because HII regions are objects of zero age when compared with the age of the Galaxy, their elemental abundances can be considered typical of primordial conditions.

The ^{7}Li isotope is observed at the surface of the oldest stars. Since the age of stars can be judged by the presence of metals, the constancy of this isotope has been interpreted as being representative of the primordial abundance.

The strongest constraint on the baryonic density comes from the primordial deuterium abundance. Ultraviolet light with a continuous flat spectrum emitted from objects at distances of $z \approx 2$–3.5 will be seen redshifted into the red range of the visible spectrum. Photoelectric absorption in intervening hydrogen along the line of sight then causes a sharp cut-off at $\lambda = 91.2$ nm, the *Lyman limit*. This can be used to select objects of a given type, which indeed are star-forming

galaxies. Deuterium is observed as a Lyman-α feature in the absorption spectra of high-redshift quasars. A recent analysis [3] gives

$$\Omega_{\rm b}(^2{\rm H})h^2 = 0.020 \pm 0.001, \tag{5.105}$$

which is more precise than any other determination. The information from the other light nucleids are in good agreement. The values of η and $\Omega_{\rm b}$ in Table A.6 come from a combined fit to ^{2}H data, CMB and large-scale structure. We defer that discussion to Section 8.4.

In Figure 5.9 the history of the Universe is summarized in nomograms relating the scales of temperature, energy, size, density and time [3]. Note that so far we have only covered the events which occurred between 10^{11} K and 10^3 K.

Nuclear synthesis also goes on inside stars where the gravitational contraction increases the pressure and temperature so that the fusion process does not stop with helium. Our Sun is burning hydrogen to helium, which lasts about 10^{10} yr, a time span which is very dependent on the mass of the star. After that, helium burns to carbon in typically 10^6 yr, carbon to oxygen and neon in 10^4 yr, those to silicon in 10 yr, and silicon to iron in 10 h, whereafter the fusion chain stops. The heavier elements have to be synthesized much later in supernova explosions, and all elements heavier than lithium have to be distributed into the intergalactic medium within the first billion years.

To sum up, Big Bang cosmology makes some very important predictions. The Universe today should still be filled with freely streaming primordial photon (and neutrino) radiation with a blackbody spectrum (5.3) of temperature related to the age of the Universe and a polarization correlated to the temperature. This relic CMB radiation (as well as the relic neutrino radiation) should be essentially isotropic since it originated in the now spherical shell of the LSS. In particular, it should be uncorrelated to the radiation from foreground sources of later date, such as our Galaxy. In Chapter 8 we shall see that these predictions have been verified for the photons (but not yet for the neutrinos).

A very important conclusion from BBN is that the Universe contains surprisingly little baryonic matter! Either the Universe is then indeed open, or there must exist other types of nonbaryonic, gravitating matter.

Problems

1. Show that an expansion by a factor a leaves the blackbody spectrum (5.3) unchanged, except that T decreases to T/a.
2. Show that the quantity $Q^2 + U^2$ is an invariant under the rotation of an angle ϕ in the (x, y)-plane, where Q and U are the Stokes parameters defined in Equation (5.8).
3. Use the definition of entropy in Equation (5.10) and the law of conservation of energy, Equation (4.24), to show what functional forms of equations of state lead to conservation of entropy.

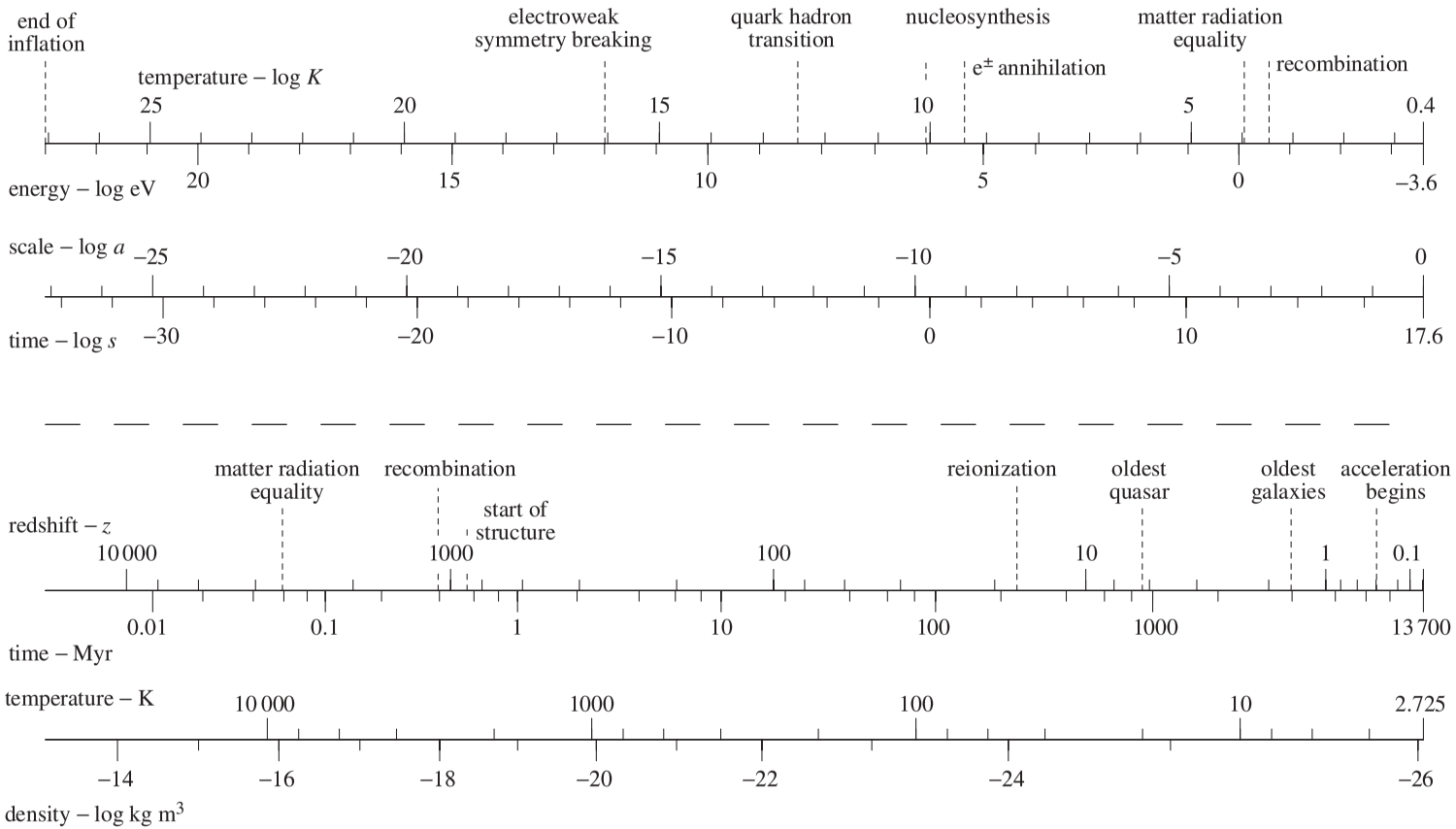


Figure 5.9 The 'complete' history of the Universe. The scales above the dashed line cover the time from the end of inflation to the present, those below from the end of radiation dominance to the present. Input constants: $H_0 = 0.71$ km s^{-1} Mpc^{-1}, $\Omega_\lambda = 0.73$, $T_0 = 2.275$ K, $t_0 = 13.7$ Gyr, and g_*, g_{*S} from Figure 5.6.

4. The flow of total energy received on Earth from the Sun is expressed by the *solar constant* 1.36×10^3 J m^{-2} s. Use Equation (5.47) to determine the surface temperature of the Sun,

$$b = \lambda T. \tag{5.106}$$

   Using this temperature and the knowledge that the dominant colour of the Sun is yellow with a wavelength of $\lambda = 0.503$ μm. What energy density does that flow correspond to?

5. The random velocity of galaxies is roughly 100 km s^{-1}, and their number density is 0.0029 per cubic megaparsec. If the average mass of a galaxy is 3×10^{44} g, what is the pressure of a gas of galaxies? What is the temperature [6]?

6. A line in the spectrum of hydrogen has frequency $\nu = 2.5 \times 10^{15}$ Hz. If this radiation is emitted by hydrogen on the surface of a star where the temperature is 6000 K, what is the Doppler broadening [6]?

7. A spherical satellite of radius r painted black, travels around the Sun at a distance d from the centre. The Sun radiates as a blackbody at a temperature of 6000 K. If the Sun subtends an angle of θ radians as seen from the satellite (with $\theta \ll 1$), find an expression for the equilibrium temperature of the satellite in terms of θ. To proceed, calculate the energy absorbed by the satellite, and the energy radiated per unit time [6].

8. Use the laws of conservation of energy and momentum and the equation of relativistic kinematics (2.69) to show that positronium cannot decay into a single photon.

9. Use the equation of relativistic kinematics (2.69) to calculate the energy and velocity of the muon from the decay (5.57) of a pion at rest. The neutrino can be considered massless.

10. What are the possible decay modes of the τ^- lepton?

11. Calculate the energy density represented by the mass of all the electrons in the Universe at the time of photon reheating when the kinetic energy of electrons is 0.2 MeV.

12. When the pions disappear below 140 MeV because of annihilation and decay, some reheating of the remaining particles occurs due to entropy conservation. Calculate the temperature-increase factor.

13. Use the equation of relativistic kinematics (2.69) and the conservation of four-momentum to calculate the energy of the photon liberated in Equation (5.86), assuming that the ^{4}He nucleus is produced at rest. (That is, $v_\mathrm{p} = v_\mathrm{t} = v_\mathrm{He} = 0$.)

14. Free nucleons are favoured over deuterons down to a radiation energy of 0.07 MeV. What is the ratio of photons with energies exceeding the deuteron binding energy 2.22 MeV to the number of protons at 0.07 MeV?

15. Propose a two-stage fusion process leading to the production of ^{12}C.
16. Gamow's penetration factor (5.99) gives a rough idea about the ignition temperatures in stellar interiors for each fusion reaction. Estimate these under the simplifying assumption that the burning rates during the different periods are inversely proportional to the time spans (given at the end of this chapter). Take the hydrogen burning temperature to be 10^4 K.

Chapter Bibliography

[1] Peacock, J. A. 1999 *Cosmological physics*. Cambridge University Press, Cambridge.
[2] Hagiwara, K. *et al*. 2002 *Phys. Rev.* D **66**, 010001-1.
[3] Burles, S., Nollett, K. M. and Turner, M. S. 2001 *Astrophys. J.* **552**, L1.
[4] Coleman, T. S. and Roos, M. 2003 *Phys. Rev.* D **68**, 027702.
[5] Bennett, C. L. *et al*. 2003 Preprint arXiv, astro-ph/0302207 and 2003 *Astrophys. J.* (In press.) and companion papers cited therein.
[6] Gasiorowicz, S. 1979 *The structure of matter*. Addison-Wesley, Reading, MA.

6

Particles and Symmetries

The laws of physics distinguish between several kinds of forces or interactions: gravitational, electroweak and strong. Although gravitation is the weakest, manifested by the fact that it takes bodies of astronomical size to make the gravitational interaction noticeable, gravitation is the most important force for understanding the Universe on a large scale. The electromagnetic force has an infinite range, like gravitation, but all astronomical objects are electrically neutral and we detect no measurable magnetic field. The weak interaction has a range of only 10^{-19} m and the strong interaction about 10^{-15} m so they are important for particles on atomic scales, but not for astronomical bodies.

The electromagnetic and weak interactions were formerly thought to be distinct but are now understood to be united as the electroweak interaction, just as happened earlier with the electric and magnetic interactions. At energies much less than 100 GeV it is convenient to distinguish between the electromagnetic and weak interactions although they are simply different manifestations of the electroweak force.

Prior to the epoch when electroweak reactions began to dominate, particles and interactions different from those we have met so far dominated the Universe. During the different epochs or phases the interactions of the particles were characterized by various symmetries governing their interactions. At each phase transition the symmetries and the physics changed radically. Thus we can only understand the electroweak epoch if we know where it came from and why. We must therefore start a journey backwards in time, where the uncertainties increase at each phase transition.

A very important symmetry is SU(2), which one usually meets for the first time in the context of electron spin. Even if electron spin is not an end to this chapter,

Introduction to Cosmology Third Edition by Matts Roos
© 2003 John Wiley & Sons, Ltd ISBN 0 470 84909 6 (cased) ISBN 0 470 84910 X (pbk)

it is a good and perhaps familiar introduction to SU(2). Thus we shall devote Section 6.1 to an elementary introduction to spin space, without the intention of actually carrying out spinor calculations.

Armed with SU(2) algebra, we study three cases of SU(2) symmetry in particle physics: the isospin symmetry of the nucleons and the weak-isospin symmetry of the leptons in Section 6.2, and the weak-isospin symmetry of the quarks in Section 6.3. We are then ready for the colour degree of freedom of quarks and gluons and the corresponding colour symmetry $\mathrm{SU}(3)_\mathrm{c}$. This leads up to the 'standard model' of particle physics, which exhibits $\mathrm{SU}(3)_\mathrm{c} \otimes \mathrm{SU}(2)_\mathrm{W} \otimes \mathrm{U}(1)_{B-L}$ symmetry.

In Section 6.4 we study the discrete symmetries of parity P, charge conjugation C and time-reversal invariance T.

In our present matter-dominated world, all the above symmetries are more or less broken, with exception only for the colour symmetry $\mathrm{SU}(3)_\mathrm{c}$ and the combined discrete symmetry CPT. This does not mean that symmetries are unimportant, rather that the mechanisms of spontaneous symmetry breaking deserve attention. We take care of this in Section 6.5.

In Section 6.6 we assemble all our knowledge of particle symmetries and their spontaneous breaking in an attempt to arrive at a grand unification theory (GUT) which unites the electroweak and the strong forces. It is a dream of physics to unite these forces as well as gravitation into a theory of everything (TOE), but their properties are so different that this has not yet succeeded.

Baryons and anti-baryons were produced from quarks in a phase transition near 200 MeV. But the reason for the baryon–anti-baryon asymmetry must be traced back to a much earlier, and not very well understood, irreversible process, in which GUT leptoquarks decayed violating baryon- and lepton-number conservation and CP. We discuss this in Section 6.7.

6.1 Spin Space

In Chapter 5 we introduced the quantal concept of spin. We found that the number n_spin of spin states of a particle was one factor contributing to its degrees of freedom g in thermal equilibrium (see Equation (5.40)). Thus horizontally and vertically polarized photons were counted as two effectively different particle species, although their physical properties were otherwise identical.

Electron Spin. Charged particles with spin exhibit magnetic moment. In a magnetic field free electrons orient themselves as if they were little magnets. The classical picture that comes to mind is an electric charge spinning around an axis parallel to the external magnetic field. Thus the resulting magnetic moment would point parallel or antiparallel to the external field depending on whether the charge is spinning right- or left-handedly with respect to the field. Although it must be emphasized that spin is an entirely quantal effect and not at all due to a spinning charge, the classical picture is helpful because the magnetic moment of

the electron—whatever the mechanism for its generation—couples to the external field just like a classical magnet.

Let us take the spin axis of the electron to be given by the *spin vector* $\boldsymbol{\sigma}$ with components σ_x, σ_y and σ_z, and let the external magnetic field $\boldsymbol{B}$ be oriented in the z direction,

$$\boldsymbol{B} = (0, 0, B_z). \tag{6.1}$$

The potential energy due to the magnetic coupling of the electron to the external field is then the *Hamiltonian operator*

$$H = A\boldsymbol{\sigma} \cdot \boldsymbol{B} = A\sigma_z B_z, \tag{6.2}$$

where A is a constant.

Measurements of H show that σ_z is *quantized*, and that it can have only two values. With a suitable choice of units, these values are ± 1. This fits the classical picture insofar as the opposite signs would correspond to right-handedly and left-handedly spinning charges, respectively. But the classical picture does not lead to quantized values: it permits a continuum of values. One consequence of the quantum dichotomy is that free electrons in thermal equilibrium with radiation each contribute $n_{\text{spin}} = 2$ degrees of freedom to g.

The above conclusions follow when the external magnetic field $\boldsymbol{B}$ was turned on. What if $\boldsymbol{B} = \boldsymbol{0}$: where do the electron spin vectors $\boldsymbol{\sigma}$ point then? The answer of course is 'anywhere', but we cannot confirm this experimentally, because we need a precise nonvanishing external field to measure H or σ_z. Thus quantum mechanics leads to a paradoxical situation: even if electrons which are not subject to observation or magnetic influence may have their spins arbitrarily oriented; by observing them we force them into either one of the two states $\sigma_z = \pm 1$.

Quantum mechanics resolves this paradox by introducing statistical laws which govern averages of ensembles of particles, but which say nothing about the individual events. Consider the free electron *before* the field B_z has been switched on. Then it can be described as having the probability P to yield the σ_z value 1 *after* the field has been switched on, and the probability $1 - P$ to yield a value of -1. The value of P is anywhere between 0 and 1, as the definition of probability requires.

After a measurement has yielded the value $\sigma_z = 1$, a subsequent measurement has a probability of 1 to yield 1 and 0 to yield -1. Thus a measurement changes the *spin state* from indefinite to definite. The definite spin states are characterized by σ_z being ± 1; the indefinite state is a *superposition* of these two states. This can be formulated either geometrically or algebraically. Let us turn to the geometrical formulation first.

Spinor Algebra. Consider an abstract two-dimensional space with orthogonal axes labelled $\boldsymbol{\chi}_+$ and $\boldsymbol{\chi}_-$, see Figure 6.1. Let us draw an arc through the points $\boldsymbol{\chi}_+ = 1$ and $\boldsymbol{\chi}_- = 1$. The length of the radius vector $\boldsymbol{\chi}$ is then unity for arbitrary polar angles ϕ, and its coordinates are $(\cos\phi, \sin\phi)$. Let us identify the square of the $\boldsymbol{\chi}_+$ coordinate, $\cos^2\phi$, with the probability P. It then follows that $1 - P = \sin^2\phi$.

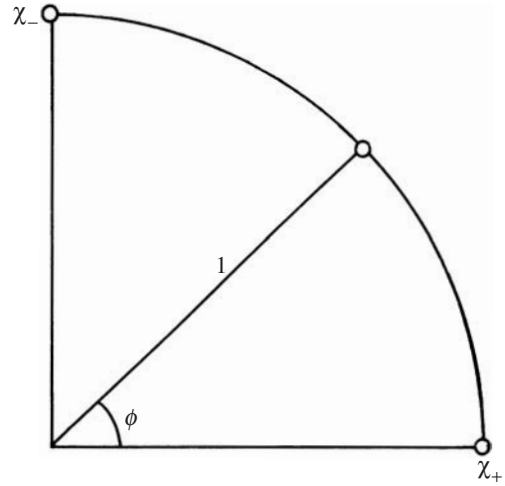


Figure 6.1 Two-component spin space.

We now identify every possible spin state with a vector from the origin to a point on the arc in Figure 6.1. These vectors in spin space are called *spinors* to distinguish them from the spin vector $\boldsymbol{\sigma}$ in ordinary space. Let us write the $\boldsymbol{\chi}_+$ and $\boldsymbol{\chi}_-$ coordinates in column form. Then the points

$$\chi_+ = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \chi_- = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{6.3}$$

are spinors corresponding to $\sigma_z = 1$ and $\sigma_z = -1$, respectively. An arbitrary point on the arc with coordinates $\cos\phi, \sin\phi$ corresponds to a linear superposition of the $\sigma_z = 1$ and $\sigma_z = -1$ states. Using the spinors (6.3) as base vectors, this can clearly be written

$$\boldsymbol{\chi} = \cos\phi \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \sin\phi \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{6.4}$$

To summarize, points on the arc correspond to states of the electron before any spin measurement has been done, and the points (6.3) correspond to *prepared states* after a measurement. Points elsewhere in the (χ_+, χ_-)-plane have no physical interpretation. The coordinates of points on the arc have no direct physical meaning, but their squares correspond to the probabilities of the outcome of a subsequent spin measurement.

The spin space has as many dimensions as there are possible outcomes: two in the electron case. It is an abstract space spanned by spinors $\boldsymbol{\chi}$, not by the spin vector $\boldsymbol{\sigma}$ of ordinary three-dimensional space. The spinor formalism allows the two spin states of the electron to be treated symmetrically: the electron is just a single two-component object, and both components have identical properties in all other respects.

Let us next turn to the algebra of spin states. The rotation of a spinor $\boldsymbol{\chi}_1$ into another spinor $\boldsymbol{\chi}_2$ corresponds algebraically to an operator U operating on $\boldsymbol{\chi}_1$,

$$\boldsymbol{\chi}_2 = \mathrm{U}\,\boldsymbol{\chi}_1. \tag{6.5}$$

Both components of $\boldsymbol{\chi}_1$ are then transformed by U, thus U in component notation must be described by a 2×2 matrix. This is most easily exemplified by the matrix σ_+, which rotates $\boldsymbol{\chi}_-$ into $\boldsymbol{\chi}_+$,

$$\sigma_+ \boldsymbol{\chi}_- = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \boldsymbol{\chi}_+, \tag{6.6}$$

and the matrix σ_-, which does the opposite. The nonunitary operators σ_+ and σ_- are called *raising* and *lowering operators*, respectively.

We noted above that the weights $\cos\phi$ and $\sin\phi$ in Equation (6.4) have no direct physical interpretation, but that their squares are real numbers with values between zero and unity. We may as well abandon the geometrical idea that the weights are real numbers, since no physical argument requires this. Let us therefore replace them by the complex numbers $a_+ = a\mathrm{e}^{\mathrm{i}\alpha}$ and $a_- = b\mathrm{e}^{\mathrm{i}\beta}$. The magnitudes a, b are real numbers which can be required to obey the same relations as $\cos\phi$ and $\sin\phi$,

$$|a_+|^2 + |a_-|^2 = a^2 + b^2 = 1, \quad 0 \leqslant |a_+| \leqslant 1, \quad 0 \leqslant |a_-| \leqslant 1, \tag{6.7}$$

but the phase angles α and β need not have any physical interpretation. Thus we can make the identifications

$$P = |a_+|^2, \qquad 1 - P = |a_-|^2. \tag{6.8}$$

The change from real weights in Equation (6.4) to complex weights does not at first sight change the physics, but it adds some useful freedom to the theory. Quantum mechanics embodies this in the important *principle of superposition*, which states that *if the spinors $\boldsymbol{\chi}_+$ and $\boldsymbol{\chi}_-$ describe physical states, then every linear superposition of them with complex coefficients $a_\pm$,*

$$\boldsymbol{\chi} = a_+ \boldsymbol{\chi}_+ + a_- \boldsymbol{\chi}_-, \tag{6.9}$$

also describes a physical state.

Unitary Transformations. It follows from the complexity of the $a_\pm$ that the matrix U in Equation (6.5) is also complex. Moreover, U is restricted to transform a point on the unit circle in Figure 6.1 into another point on the circle. One then proves easily that the operator U must be *unitary*, obeying

$$\mathrm{U}\,\mathrm{U}^\dagger = \mathrm{U}^\dagger\,\mathrm{U} = 1, \tag{6.10}$$

where the superscript '†' implies transposition and complex conjugation. Actually, the unitarity condition (6.10) follows from Equations (6.5) and (6.7).

If the spin space were one-dimensional the unitarity condition (6.10) would be satisfied by any pure *phase transformation*

$$\mathrm{U} = \mathrm{e}^{\mathrm{i}\theta \cdot 1}. \tag{6.11}$$

The number '1' is of course superfluous in the exponent, but we keep it for later reference. All operators of this form are elements of a mathematical *group* called

U(1). Here U stands for unitary and (1) for the *order* of the group. The phase angle θ is a *real parameter of the group* having a global value all over space-time, that is, it does not depend on the space-time coordinates. The transformation (6.11) is therefore called a *global gauge transformation*.

A similar situation occurs in another familiar context: Maxwell's equations for the electromagnetic field are invariant under a phase transformation

$$\mathrm{U} = \mathrm{e}^{\mathrm{i}\,Q\theta(x)}, \tag{6.12}$$

where the electric charge Q is a conserved quantity. Now, however, the parameter of the group is the product $Q\theta(x)$, which depends on the local space-time coordinate x through the function $\theta(x)$. The U(1) symmetry implies that Maxwell's equations are independent of the local choice of $\theta(x)$. The transformation (6.12) is then called a *local gauge transformation* (see, for example, [1]). Because this gauge symmetry is exact, the *gauge boson* of the theory, the photon, has exactly zero mass.

This situation is quite similar to the principle of covariance, which demanded that the laws of physics should be independent of the choice of local space-time coordinates. The principle of covariance is now replaced by the *gauge principle*, which applies to gauge field theories.

In the spin space of the electron, the operators U are represented by unitary 2×2 matrices which, in addition, obey the special condition

$$\det U = 1. \tag{6.13}$$

This defines them to be elements of the global group SU(2), of order two. The letter S in SU(2) stands for 'special' condition. In group theory parlance, the two-component spinors in Equation (6.3) are *doublet representations* of SU(2).

It is possible to express the U operators in terms of exponentiations,

$$\mathrm{U} = \mathrm{e}^{\mathrm{i}H}, \tag{6.14}$$

analogously to the one-dimensional case (6.11). This requires the quantities H to be complex 2×2 matrices as well. Substituting the expression (6.14) into the unitarity condition (6.10), we have

$$\mathrm{U}\,\mathrm{U}^{\dagger} = \mathrm{e}^{\mathrm{i}(H - H^{\dagger})} = 1. \tag{6.15}$$

It follows that the operators H must be *Hermitian*,

$$H = H^{\dagger}. \tag{6.16}$$

Moreover, the special condition (6.13) requires the H matrices to be traceless.

Pauli Matrices. The most general way to write a 2×2 Hermitian matrix is

$$H = \theta_1 \sigma_x + \theta_2 \sigma_y + \theta_3 \sigma_z, \tag{6.17}$$

where the σ_i are the *Pauli matrices* invented by *Wolfgang Pauli* (1900–1958)

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \sigma_y = \begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix}, \qquad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{6.18}$$

Note that all the σ_i are traceless, and only σ_z is diagonal.

Comparing the exponent in Equation (6.11) with the expression (6.17) we see that the single parameter θ in the one-dimensional case corresponds to three real parameters $\theta_1, \theta_2, \theta_3$ in SU(2). The superfluous number 1 in U(1) is the vestige of the Pauli matrices appearing in SU(2). It shares with them the property of having the square 1.

The number 1 generates the ordinary algebra—a trivial statement, indeed!—whereas the Pauli matrices generate a new, *noncommutative algebra*. In *commutative* or *Abelian algebras* the product of two elements θ_1 and θ_2 can be formed in either order:

$$\theta_1\theta_2 - \theta_2\theta_1 \equiv [\theta_1, \theta_2] = 0.$$

Here the square-bracketed expression is called a *commutator*. In the *non-Abelian* algebra SU(2) the commutator of two elements does not in general vanish. For instance, the commutator of two Pauli matrices σ_i is

$$[\sigma_i, \sigma_j] = 2i\sigma_k, \tag{6.19}$$

where i, j, k represent any cyclic permutation of x, y, z.

In quantum theory all *possible observations* are represented algebraically by *linear operators*, operating on *physical states*. When the states are described by spinors as in Equation (6.3), the operators are *diagonal matrices*. The *possible outcomes* of the observation are the values on the diagonal.

For the case of an electron in the *spin state* $\boldsymbol{\chi}_\pm$, the relation between the *operator* σ_z, describing the observation of spin in the z direction, the outcome ± 1, and the state $\chi_\pm$ is formulated

$$\sigma_z\chi_+ = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \chi_+, \tag{6.20}$$

$$\sigma_z\chi_- = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} 0 \\ 1 \end{pmatrix} = -\begin{pmatrix} 0 \\ 1 \end{pmatrix} = -\chi_-. \tag{6.21}$$

These two equations are called *eigenvalue equations*. The $\boldsymbol{\chi}_\pm$ are said to be *eigenstates* of the operator σ_z, with the *eigenvalues* ± 1, respectively. One can always choose a base where one (but only one) of the operators σ_x, σ_y or σ_z is diagonal. Note that the general spinor (6.9) is not an eigenstate of σ_z because

$$\sigma_z\boldsymbol{\chi} = \sigma_z(a_+\boldsymbol{\chi}_+ + a_-\boldsymbol{\chi}_-) = (a_+\boldsymbol{\chi}_+ - a_-\boldsymbol{\chi}_-) \neq \pm\boldsymbol{\chi}.$$

The important lesson of this is that *possible observations are operators* which can be represented by *diagonal matrices*, and the numbers appearing on the *diagonal* are the *possible outcomes* of the observation. Moreover, the operators must be linear, since they operate in a linear space, transforming spinors into spinors.

In ordinary space the spin vector $\boldsymbol{\sigma}$ has a length which of course is a real positive number. Since its projection on the z-axis is either $\sigma_z = +\frac{1}{2}$ or $\sigma_z = -\frac{1}{2}$, the length of $\boldsymbol{\sigma}$ must be $\sigma \equiv |\boldsymbol{\sigma}| = \frac{1}{2}$. The sign of σ_z indicates that $\boldsymbol{\sigma}$ is parallel or antiparallel to the z direction. Consider a system formed by two electrons a and b with spin vectors $\boldsymbol{\sigma}_\mathrm{a}$ and $\boldsymbol{\sigma}_\mathrm{b}$ and spinor states χ_+^a, χ_-^a, χ_+^b and χ_-^b. The sum vector

$$\boldsymbol{\sigma} = \boldsymbol{\sigma}_\mathrm{a} + \boldsymbol{\sigma}_\mathrm{b}$$

can clearly take any values in the continuum between zero and unity, depending on the relative orientations. However, quantum mechanics requires $\boldsymbol{\sigma}$ to be quantized to integral values, in this case to 0 or 1. This has important consequences for atomic spectroscopy and particle physics.

6.2 SU(2) Symmetries

The proton and the neutron are very similar, except in their electromagnetic properties: they have different charges and magnetic moments. Suppose one could switch off the electromagnetic interaction, leaving only the *strong interaction* at play. Then the p and n fields would be identical, except for their very small mass difference. Even that mass difference one could explain as an electromagnetic effect, because it is of the expected order of magnitude.

Nucleon Isospin. Making use of SU(2) algebra one would then treat the nucleon N as a two-component state in an abstract *charge space*,

$$\boldsymbol{N} = \begin{pmatrix} p \\ n \end{pmatrix}. \tag{6.22}$$

The p and n fields are the base vectors spanning this space,

$$\boldsymbol{p} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \boldsymbol{n} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{6.23}$$

In analogy with the spin case, these states are the eigenstates of an operator I_3, completely unrelated to spin, but having the algebraic form of $\frac{1}{2}\sigma_z$ and eigenvalues

$$I_3 = \pm\tfrac{1}{2}. \tag{6.24}$$

Thus the proton with charge $Q = +1$ has $I_3 = +\frac{1}{2}$, and the neutron with $Q = 0$ has $I_3 = -\frac{1}{2}$. It is then convenient to give I_3 physical meaning by relating it to charge,

$$Q = \tfrac{1}{2} + I_3. \tag{6.25}$$

It follows that the *charge operator* in matrix form is

$$Q = \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \tag{6.26}$$

where the charges of p and n, respectively, appear on the diagonal.

One can also define two operators $I_1 = \frac{1}{2}\sigma_x$ and $I_2 = \frac{1}{2}\sigma_y$ in order to recover the complete SU(2) algebra in the space spanned by p and n. These operators interchange the charge states, for instance

$$I_1\boldsymbol{N} = \frac{1}{2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} p \\ n \end{pmatrix} = \frac{1}{2}\begin{pmatrix} n \\ p \end{pmatrix}. \tag{6.27}$$

The I_1, I_2, I_3 are the components of the *isospin* vector $\boldsymbol{I}$ in an abstract three-dimensional space. This contrasts with the spin case where $\boldsymbol{\sigma}$ is a vector in ordinary three-dimensional space.

The advantage of this notation is that the strong interactions of protons and neutrons as well as any linear superposition of them are treated symmetrically. One says that strong interactions possess *isospin symmetry*. Just as in the case of electron spin, this is a global symmetry.

In nature, the isospin symmetry is not exact because one cannot switch off the electromagnetic interactions, as we supposed to begin with. Since the main asymmetry between the proton and the neutron is expressed precisely by their different electric charges, electromagnetic interactions are not isospin symmetric. However, the strong interactions are so much stronger, as is witnessed by the fact that atomic nuclei containing large numbers of protons do not blow apart in spite of their mutual Coulomb repulsion. Thus isospin symmetry is approximate, and it turns out to be a more useful tool in particle physics than in nuclear physics.

Lepton Weak Isospin. The lessons learned from spin space and isospin symmetry have turned out to be useful in still other contexts. As we have seen in Section 5.3, the electron and its neutrino form a family characterized by the conserved e-lepton number L_e. They participate in similar electroweak reactions, but they differ in mass and in their electromagnetic properties, as do the proton and neutron. The neutrinos are very light, so the mass difference is too important to be blamed on their different electric charges. Quite distinctly from the isospin symmetry, which is at best useful, there is another SU(2) symmetry which is of fundamental importance: the leptons are considered to be components of three SU(2) doublets,

$$\boldsymbol{\ell}_e = \begin{pmatrix} \nu_e \\ e^- \end{pmatrix}, \qquad \boldsymbol{\ell}_\mu = \begin{pmatrix} \nu_\mu \\ \mu^- \end{pmatrix}, \qquad \boldsymbol{\ell}_\tau = \begin{pmatrix} \nu_\tau \\ \tau^- \end{pmatrix}. \tag{6.28}$$

Their antiparticles form three similar doublets where the up/down order is reversed. Thus we meet yet another abstract space, *weak-isospin space*, which is not identical to either spin space or isospin space; only the algebraic structure is the same. This space is spanned by spinor-like base vectors

$$\boldsymbol{\nu}_e = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad e^- = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{6.29}$$

Making use of the notation (6.28) the reactions (5.31), (5.36), and (5.37) can be written as one reaction,

$$\overset{(-)}{\ell_i} + \overset{(-)}{\ell_j} \longrightarrow \overset{(-)}{\ell_i} + \overset{(-)}{\ell_j}, \tag{6.30}$$

where the subscripts i and j refer to e, μ or τ. In analogy with the spin and isospin cases, the states (6.29) are the eigenstates of an operator

$$T_3 = \tfrac{1}{2}\sigma_z = \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{6.31}$$

The eigenvalues $T_3 = \pm\frac{1}{2}$ appearing on the diagonal are called *weak charge*. One can also define a weak-isospin vector $\boldsymbol{T}$ with components T_1, T_2, T_3 spanning an abstract three-dimensional space.

In the isospin case we had to 'switch off' electromagnetism in order to achieve a global symmetry between the differently charged p and n states of the nucleon. In the case of the electroweak theory, the particle states are represented by gauge fields which are locally gauged (see, for example, [1]). In addition, a trick has been invented which incorporates both electric charge and weak charge in the symmetry group. This trick is to enlarge the gauge symmetry to the direct product of two local gauge symmetry groups,

$$\mathrm{SU}(2)_{\mathrm{w}} \otimes \mathrm{U}(1)_Y.$$

Here w stands for weak isospin, and Y is a new quantum number called *weak hypercharge*, the parameter of a $\mathrm{U}(1)$ group. If one defines the latter by

$$\tfrac{1}{2}Y = Q - T_3, \tag{6.32}$$

all the leptons have $Y = -1$ regardless of charge, and all the anti-leptons have $Y = 1$.

The assumption that nature observes $\mathrm{SU}(2)_{\mathrm{w}} \otimes \mathrm{U}(1)_Y$ symmetry implies that the electroweak interaction does not see any difference between the neutrino and the electron fields and linear superpositions of them,

$$\ell_{\mathrm{e}} = a_+ \nu_{\mathrm{e}} + a_- \mathrm{e}^-;$$

they all have the same weak hypercharge and the same L_{e}. That the symmetry is a local gauge symmetry implies that the laws of electroweak interactions are independent of the 'local' choice of gauge functions, analogous to $\theta(x)$ in Equation (6.12). However, nature does not realize this symmetry exactly; it is a *broken symmetry*. This is seen by the fact that of the four gauge bosons γ, Z^0, W^+, W^- mediating the electroweak interaction, three are massive. In an exact local gauge-symmetric theory all gauge bosons are massless, as is the photon in the case of the exact $\mathrm{U}(1)$.

Actually, the electroweak force is the outcome of a long and difficult search to unify the forces of nature. The first milestone on this road was set up by Maxwell when he managed to unify electricity and magnetism into one electromagnetic force. Since then science has been concerned with four forces of nature: the strong force responsible for the stability of nuclei, the electromagnetic force responsible for atomic structure and chemistry, the weak force which played such an important cosmological role during the late radiation era and the gravitational force acting during the matter-dominated era. The latter three forces are described by local gauge-field theories.

Einstein attempted in vain to unify gravitation and electromagnetism during his last 20 years. A breakthrough came in 1967 when *Sheldon Glashow*, *Steven Weinberg* and *Abdus Salam* succeeded in unifying the weak and electromagnetic forces into the electroweak interaction. Since then the goal has been to achieve a *grand unified theory* (GUT), which would unify strong and electroweak interactions, and ultimately gravitation as well in a *theory of everything* (TOE).

6.3 Hadrons and Quarks

The spectrum of *hadrons* participating in strong interactions is extremely rich. The hadrons comprise two large classes of particles already encountered, the baryons and the mesons. Charged hadrons also have electromagnetic interactions, and all hadrons have weak interactions, but strong interactions dominate whenever possible.

The reaction rate of strong interactions exceeds by far the rate of other interactions. Weak decays like reactions (5.57) and (5.59) typically require mean lives ranging from microseconds to picoseconds. An electromagnetic decay like

$$\pi^0 \longrightarrow \gamma + \gamma$$

takes place in less than 10^{-16} s, and heavier particles may decay 1000 times faster. Strongly decaying hadrons, however, have mean lives of the order of 10^{-23} s.

Some simplification of the hadron spectrum may be achieved by introducing isospin symmetry. Then the nucleon N stands for n and p, the pion π stands for π^+, π^0, π^-, the *kaon* K for the *strange mesons* K^+, K^0 with mass 495 MeV, etc.

Quarks. In 1962 *Murray Gell-Mann* and *George Zweig* realized that the hadron spectrum possessed more symmetry than isospin symmetry. They proposed that all hadrons known could be built out of three hypothetical states called *quarks*, $q = \mathrm{u}, \mathrm{d}, \mathrm{s}$, which spanned a three-dimensional space with SU(3) gauge symmetry. This SU(3) group is an extension of the isospin SU(2) group to include *strangeness*, an additive quantum number possessed by the kaon and many other hadrons. The *up* and *down* quark fields u, d form an isospin doublet like Equation (6.22), whereas the *strange quark* s is an isospin-neutral singlet. Together they form the basic building block of SU(3), a triplet of quarks of three *flavours*. The quarks are fermion fields just like the leptons, and in spin space they are SU(2) doublets.

In the quark model the mesons are $\mathrm{q}\bar{\mathrm{q}}$ bound states, and the baryons are qqq bound states. The differences in hadron properties can be accounted for in two ways: the quarks can be excited to higher angular momenta, and in addition the three flavours can enter in various combinations. For instance, the nucleon states are the ground state configurations

$$p = \mathrm{uud}, \qquad n = \mathrm{udd}. \tag{6.33}$$

The mesonic ground states are the pions and kaons with the configurations

$$\left.\begin{aligned} \pi^+ = \mathrm{u}\bar{\mathrm{d}}, \qquad \pi^0 = \tfrac{1}{\sqrt{2}}(\mathrm{u}\bar{\mathrm{u}} - \mathrm{d}\bar{\mathrm{d}}), \qquad \pi^- = \mathrm{d}\bar{\mathrm{u}}, \\ K^+ = \mathrm{u}\bar{\mathrm{s}}, \qquad K^0 = \mathrm{d}\bar{\mathrm{s}}, \end{aligned}\right\} \tag{6.34}$$

as well as the η and η' which are linear combinations of $\mathrm{u}\bar{\mathrm{u}}$, $\mathrm{d}\bar{\mathrm{d}}$ and $\mathrm{s}\bar{\mathrm{s}}$.

After the discovery in the 1960s of the electroweak gauge symmetry $\mathrm{SU}(2)_\mathrm{w} \otimes \mathrm{U}(1)_Y$ for the then known e- and μ-lepton families and the u, d family of quarks,

it was realized that this could be the fundamental theory of electroweak interactions. But the theory clearly needed a fourth flavour quark to complete a second $SU(2)_w$ doublet together with the s-quark. To keep the s-quark as a singlet would not do: it would be T_3-neutral and not feel the weak interactions. In 1974 the long predicted *charmed quark* c was discovered simultaneously by two teams, an MIT team led by *Sam Ting*, and a SLAC team lead by *Burt Richter*.

The following year the issue was confused once more when another SLAC team, led by *Martin Perl*, discovered the τ lepton, showing that the lepton families were three. This triggered a search for the τ neutrino and the corresponding two quarks, if they existed. In 1977 a team at the Cornell e^+e^- collider CESAR led by *Leon Lederman* found the fifth quark with the same charge as the d and s, but with its own new flavour. It was therefore a candidate for the bottom position in the third quark doublet. Some physicists lacking the imagination of those who invented 'strangeness' and 'charm', baptized it *bottom quark* b, although the name *beauty* has also been used. The missing companion to the bottom quark in the third $SU(2)_w$ doublet was prosaically called *top quark*, t. The fields of the three quark families can then be ordered as

$$\begin{pmatrix} u \\ d \end{pmatrix}, \qquad \begin{pmatrix} c \\ s \end{pmatrix}, \qquad \begin{pmatrix} t \\ b \end{pmatrix}. \tag{6.35}$$

The top quark was discovered by the CDF and D0 teams at the Fermilab in 1994–1995. The known ground state charm and bottom mesons are, respectively,

$$D^+ = c\bar{d}, \qquad D^0 = c\bar{u}, \qquad D_s^+ = c\bar{s},$$
$$B^+ = u\bar{b}, \qquad B^0 = d\bar{b}, \qquad B_s^0 = s\bar{b}.$$

In addition $c\bar{c}$ and $b\bar{b}$ states are known.

The strong interaction symmetry, which had started successfully with SU(3) for three quarks, would logically be enlarged to $SU(n)$ for quarks of n flavours. However, the quark masses, although not directly measurable, are so vastly different that even SU(4) is a badly broken symmetry and not at all useful. Only the isospin SU(2) subgroup and the flavour SU(3) subgroup continue to be useful, in particular for the classification of hadrons.

It follows from the quark structure (6.33) of the nucleons that each quark possesses baryon number $B = \frac{1}{3}$. The $SU(2)_w$ symmetry requires all the $T_3 = \frac{1}{2}$ (upper) states in the doublets to have the same charge Q_u, and all the $T_3 = -\frac{1}{2}$ (lower) states to have the charge Q_d. To match the nucleon charges, the charges of the quarks have to satisfy the relations

$$2Q_u + Q_d = 1, \qquad Q_u + 2Q_d = 0. \tag{6.36}$$

The solution chosen by nature is

$$Q_u = \tfrac{2}{3}, \qquad Q_d = -\tfrac{1}{3}. \tag{6.37}$$

It now follows from the definition of weak hypercharge Y in Equation (6.32) that all the quarks have the same weak hypercharge, for instance

$$Y_d = -\tfrac{2}{3} + 1 = \tfrac{1}{3}. \tag{6.38}$$

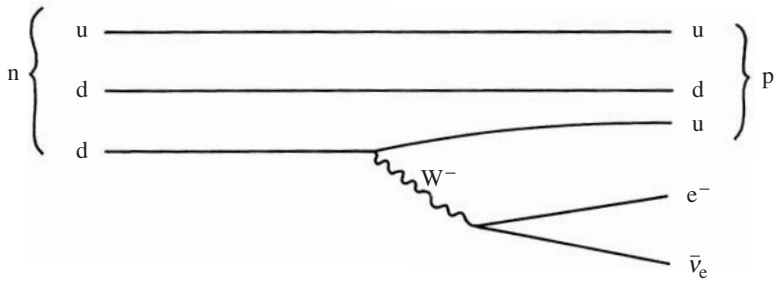


Figure 6.2 Feynman diagram for neutron β decay with quark lines.

Thus the quarks differ from the leptons in weak hypercharge. Actually, we can get rid of the somewhat artificial notion of weak hypercharge by noting that

$$Y = B - L, \tag{6.39}$$

where B is the baryon number not possessed by leptons, and L is the lepton number not possessed by baryons. Thus for leptons $B - L = -1$, whereas for quarks it is $\frac{1}{3}$.

The electroweak interactions of the leptons and quarks are mediated by the *gauge bosons* γ, Z^0, W^+, W^-. Two examples of this interaction were illustrated by the Feynman diagrams in Figures 5.2 and 5.3, where each line corresponds to a particle. Figure 6.2 shows the Feynman diagram for neutron β decay, reaction (5.96), where the decomposition into quarks is explicit, a nucleon corresponding to three quark lines. As is seen, the decay of a neutron involves the transformation of a d quark of charge $-\frac{1}{3}$ into a u quark of charge $\frac{2}{3}$ and a virtual W^- boson. The final quark system is therefore that of a proton. Two of the quarks do not participate in the reaction at all; they remain *spectators* of what is going on. Subsequently, the virtual W^- produces a lepton–anti-lepton pair, conserving electric charge and keeping the total lepton number at $L = 0$.

The strong interactions of hadrons, which were never very well understood, obviously had to be replaced by interactions at the quark level. For this a new mediator is needed, the *gluon*, which is a vector boson like the previously mentioned mediators of interactions, but massless like the photon. The gluon is also then responsible for binding quarks into hadrons, but it does not feel the leptons. The field theory describing the interactions of quarks and gluons is called *quantum chromodynamics* (QCD)—the reason for the word *chromo* will be explained next.

Colour. The quarks have another intrinsic property which they do not share with the leptons. It appears from problems in hadron spectroscopy, and from the rates of certain hadronic reactions which occur three times faster than expected, that each quark actually must come in three versions. These versions do not differ from each other in any respect encountered so far, so they require a new property called *colour*. To distinguish quarks of different colour, one may choose to call them red,

blue and yellow (R, B, Y), for instance. They span an abstract three-dimensional space with $SU(3)_c$ symmetry ('c' for colour), and base vectors

$$\boldsymbol{q}_R = \begin{pmatrix}1\\0\\0\end{pmatrix}, \qquad \boldsymbol{q}_B = \begin{pmatrix}0\\1\\0\end{pmatrix}, \qquad \boldsymbol{q}_Y = \begin{pmatrix}0\\0\\1\end{pmatrix}, \tag{6.40}$$

where q stands for the flavours u, d, c, s, t, b.

Colour is an absolutely conserved quantum number, in contrast to flavour, which is conserved in strong and electromagnetic interactions, but broken in weak interactions. Since the gluon interacts with quarks mediating the *colour force*, it must itself carry colour, so that it can change the colour of a quark. The same situation occurs in electroweak interactions where the conservation of charge, as for instance in Figure 6.2, requires the W to carry charge so that it can change a d quark into a u quark. Gluons interact with gluons because they possess colour charge, in contrast to photons, which do not interact with photons because of their lack of electric charge.

Since there are quarks of three colours, there must exist nine gluons, one for each distinct colour pair. Thus the gluon colours are $B\bar{B}$, $B\bar{R}$, $B\bar{Y}$, $R\bar{B}$, $R\bar{R}$, $R\bar{Y}$, $Y\bar{B}$, $Y\bar{R}$, $Y\bar{Y}$. One linear combination of $B\bar{B}$, $R\bar{R}$, and $Y\bar{Y}$, which is colour-neutral and totally antisymmetric under the exchange of any two colours, is a *singlet* state and the interaction is completely blind to it. Thus there are eight distinct gluons.

Since the colour property is not observed in hadrons, they must be colour-neutral. How can one construct colour-neutral compounds of coloured constituents? For the mesons which are quark–anti-quark systems, the answer is quite simple: for a given quark colour the anti-quark must have the corresponding anti-colour. Thus for instance the π^+ meson is a $u\bar{d}$ system when we account for flavours only, but if we account also for colour three π^+ mesons are possible, corresponding to $u_B\bar{d}_{\bar{B}}$, $u_R\bar{d}_{\bar{R}}$ and $u_Y\bar{d}_{\bar{Y}}$, respectively. Each of these is colour-neutral, so hadronic physics does not distinguish between them. Also the baryons which are qqq states must be colour-neutral. This is possible for a totally antisymmetric linear combination of three quarks q, q′, q″, having three colours each.

Asymptotic Freedom. It is a curious fact that free quarks never appear in the laboratory, in spite of ingenious searches. When quarks were first invented to explain the spectroscopy of hadrons, they were thought to be mere abstractions without real existence. Their reality was doubted by many people since nobody had succeeded in observing them. But it was gradually understood that their nonobservability was quite natural for deep reasons related to the properties of vacuum.

The quark and the anti-quark in a meson are like the north and south poles of a magnet which cannot be separated, because there exists nothing such as a free magnetic north pole. If one breaks the bound of the poles, each of the pieces will become a new magnet with a north and a south pole. The new opposite poles are generated at the break, out of the vacuum so to speak . Similarly, if one tries to break a $q\bar{q}$ pair, a new $q\bar{q}$ pair will be generated out of the vacuum at the break, and one thus obtains two new $q\bar{q}$ mesons which are free to fly away.

To account for this property one must assign curious features to the potential responsible for the binding of the $q\bar{q}$ system: the larger the interquark distance, the stronger the potential. Inversely, at very small interquark distances the potential can be so weak that the quarks are essentially free! This feature is called *asymptotic freedom*.

Higher Symmetries. The SU(3) algebra is a straightforward generalization of SU(2) to order three. Since the base vectors (6.40) have three components, the operators also in that space must be 3×3 matrices which generalize the Pauli matrices. Of these, two are diagonal, corresponding to two observable properties.

We are now ready to combine the $SU(3)_c$ symmetry with the electroweak symmetry in a global symmetry for the gluonic interactions of quarks and the electroweak interactions of leptons and quarks. The most elegant way would be if nature were symmetric under a larger group, say SU(5) or SO(10), which has SU(2) and SU(3) as subgroups. However, no satisfactory larger group has been found, so the global symmetry group

$$G_S \equiv SU(3)_c \otimes SU(2)_w \otimes U(1)_{B-L} \tag{6.41}$$

seems to be the less elegant direct product of the three symmetry groups. This symmetry is referred to as the *standard model* in particle physics (not to be confused with the standard model in Big Bang cosmology). This will play an important role in the discussion of the primeval Universe.

6.4 The Discrete Symmetries C, P, T

According to the cosmological principle, the laws of physics should be independent of the choice of space-time coordinates in a homogeneous and isotropic universe. Indeed, all laws governing physical systems in isolation, independent of external forces, possess *translational symmetry* under the displacement of the origin in three-space as well as in four-space. Such systems also possess *rotational symmetry* in three-space. Translations and rotations are continuous transformations in the sense that a finite transformation (a translation by a finite distance or a rotation through a finite angle) can be achieved by an infinite sequence of infinitesimal transformations.

Space Parity. A different situation is met in the transformation from a right-handed coordinate system in three-space to a left-handed one. This is achieved by reflecting the coordinate system in a point, or by replacing the x, y, z coordinates simultaneously by $-x, -y, -z$, respectively. This transformation cannot be achieved by an infinite sequence of infinitesimal transformations, and it therefore represents a *discrete transformation*.

The mirror reflection in three-space is called *parity transformation*, and the corresponding *parity operator* is denoted P. Obviously, every vector $\boldsymbol{v}$ in a right-

handed coordinate system is transformed into its negative in a left-handed coordinate system,

$$\mathrm{P}\boldsymbol{v} = -\boldsymbol{v}.$$

This has the structure of an eigenvalue equation: $\boldsymbol{v}$ is an *eigenvector* of P with the eigenvalue $P = -1$. A function $f(\boldsymbol{r})$ of the position vector $\boldsymbol{r}$ is transformed by P into

$$\mathrm{P}f(\boldsymbol{r}) = f(-\boldsymbol{r}). \tag{6.42}$$

Let us take $f(\boldsymbol{r})$ to be a scalar function which is either symmetric under the parity transformation, $f(-\boldsymbol{r}) = f(\boldsymbol{r})$, or antisymmetric, $f(-\boldsymbol{r}) = -f(\boldsymbol{r})$. In both cases Equation (6.42) is an eigenvalue equation with $f(\boldsymbol{r})$ the *eigenfunction* of P having the eigenvalue $P = +1$ or $P = -1$, respectively. Thus, scalars transform under P in two ways: those corresponding to *even parity* $P = +1$ are called (true) *scalars*, those corresponding to *odd parity* $P = -1$ are called *pseudoscalars*.

It may seem intuitively natural that the laws of physics should possess this left–right symmetry. The laws of classical mechanics in fact do, and so do Maxwell's laws of electrodynamics and Newton's and Einstein's laws of gravitation. All particles transform under P in some particular way which may be that of a scalar, a pseudoscalar, a vector or yet another. One can then consider that this is an intrinsic property, *parity* $P = \pm 1$, if the particles are eigenstates of P. The bosons are but the fermions are not eigenstates of P because of their spinor nature (recall that the W and Z are vector bosons). However, fermion–anti-fermion pairs are eigenstates with odd parity, $P = -1$. The strong interactions conserve parity, and so do the electromagnetic interactions, from the evidence of Maxwell's equations. In a parity-conserving universe there is no way to tell in an absolute sense which direction is left and which right.

It came as a surprise then when, in 1957, it was discovered that the weak interactions turned out to violate left–right symmetry, in fact maximally so. In a weak interaction the intrinsic parity of a particle could change. Thus if we communicated with a being in another galaxy we could tell her in an absolute sense which direction is left by instructing her to do a β decay experiment.

Helicity States. A consequence of this maximal violation is the *helicity* property of the leptons and quarks. For a particle with spin vector $\boldsymbol{\sigma}$ moving in some frame of reference with momentum $\boldsymbol{p}$, the helicity is defined as

$$H = \frac{\sigma \cdot \boldsymbol{p}}{|\boldsymbol{p}|}. \tag{6.43}$$

Particles with $H < 0$ are called *left handed* and particles with $H > 0$ are called *right handed*. The maximal left–right symmetry violation implies that left-handed leptons and quarks have couplings to the weak interaction gauge bosons W and Z, whereas the couplings of the right-handed ones are strongly suppressed. This is true in particular for the neutrinos which have such small masses as to make the right-handed ones practically inert. The rate for W and Z mediated scattering of right-handed neutrinos is suppressed, in comparison with the left-handed rate,

by a factor of the order of m_ν^2/E^2, where E is the energy in the centre-of-mass system, and the neutrino masses are $\langle m_\nu \rangle < 0.23$ eV. This is the reason why the neutrinos contributed only $\frac{7}{4}$ to g_* in Equation (5.52), while the charged leptons contributed $\frac{7}{2}$.

All the above is true also for anti-leptons and anti-quarks, except that their helicity has the reversed sign. Thus their weak interactions are dominantly right-handed, and left-handed anti-neutrinos are practically inert.

Charge Conjugation. Let us now introduce another discrete operator C called *charge conjugation*. The effect of C on a particle state is to turn it into its own antiparticle. For flavourless bosons like the pion this is straightforward because there is no fundamental difference between a boson and its anti-boson, only the electric charge changes, e.g.

$$\mathrm{C}\pi^+ = \pi^-. \tag{6.44}$$

Thus the charged pion is not an eigenstate of this operator, but the π^0 is. The C operator reverses the signs of all flavours, lepton numbers, and the baryon number.

Weak interactions are in fact symmetric under CP to a very good approximation. The combined operator CP is useful because it transforms left-handed leptons into right-handed anti-leptons, both of which are observed states:

$$\mathrm{CP}\ell_\mathrm{L} = \mathrm{C}\nu_\mathrm{R} = \bar{\ell}_\mathrm{R}. \tag{6.45}$$

Some reactions involving kaons exhibit a tiny CP violation, of the order of 0.22% relative to CP-conserving reactions. The decays of B-mesons, which are bu- and bd-quark systems, also exhibit a very small CP violation. It turns out that this tiny effect is of fundamental importance for cosmology, as we shall see in Section 6.7. The reason for CP violation is not yet understood.

The strong interactions violate CP with an almost equal amount but with opposite sign, such that the total violation cancels, or in any case it is less than 10^{-9}. Why this is so small is also not known.

Time Reversal. A third discrete symmetry of importance is *time reversal* T, or symmetry under inversion of the arrow of time. This is a mirror symmetry with respect to the time axis, just as parity was a mirror symmetry with respect to the space axes. All physical laws of reversible processes are formulated in such a way that the replacement of time t by $-t$ has no observable effect. The particle reactions in Section 5.3 occur at the same rate in both directions of the arrow (to show this, one still has to compensate for differences in phase space, i.e. the bookkeeping of energy in endothermic and exothermic reactions).

CPT Symmetry. Although time reversal is not very important in itself, for instance particles do not carry a conserved quantum number related to T, it is

one factor in the very important combined symmetry CPT. According to our most basic notions in theoretical physics, CPT symmetry must be absolute. It then follows from the fact that CP is not an absolute symmetry, but slightly violated, that T must be violated by an equal and opposite amount.

In a particle reaction, CPT symmetry implies that a *left-handed particle entering* the interaction region from the x-direction is equivalent to a *right-handed antiparticle leaving* the region in the x-direction. One consequence of this is that particles and antiparticles must have exactly the same mass and, if they are unstable, they must also have exactly the same mean life.

Needless to say, many ingenious experiments have been and still are carried out to test CP violation and T violation to ever higher precision. CPT symmetry will probably be tested when sufficient numbers of anti-hydrogen atoms have been produced, and their properties will be compared with those of hydrogen.

6.5 Spontaneous Symmetry Breaking

As we have seen, nature observes exactly very few symmetries. In fact, the way a symmetry is broken may be an important ingredient in the theory. As an introduction to *spontaneous breaking* of particle symmetries, let us briefly study some simple examples from other fields. For further reading, see, for example, references [1]–[4].

Classical Mechanics. Consider a cylindrical steel bar standing on one end on a solid horizontal support. A vertical downward force is applied at the other end. This system is obviously symmetrical with respect to rotations around the vertical axis of the bar. If the force is increased beyond the strength of the steel bar, it buckles in one direction or another. At that moment, the cylindrical symmetry is broken.

Bar Magnet. An iron bar magnet heated above the Curie temperature 770 °C loses its magnetization. The minimum potential energy at that temperature corresponds to a completely random orientation of the magnetic moment vectors of all the electrons, so that there is no net magnetic effect. This is shown in Figure 6.3, where the potential energy follows a parabola with its apex at zero magnetization. Since no magnetization direction is selected, this bar magnet possesses full rotational symmetry. The corresponding symmetry group is denoted $O(3)$ for orthogonal rotations in three-space.

As the bar magnet cools below a temperature of 770 °C, however, this symmetry is spontaneously broken. When an external magnetic field is applied, the electron magnetic-moment vectors align themselves, producing a net collective macroscopic magnetization. The corresponding curve of potential energy has two deeper minima symmetrically on each side of zero magnetization (see Figure 6.3). They distinguish themselves by having the north and south poles reversed. Thus

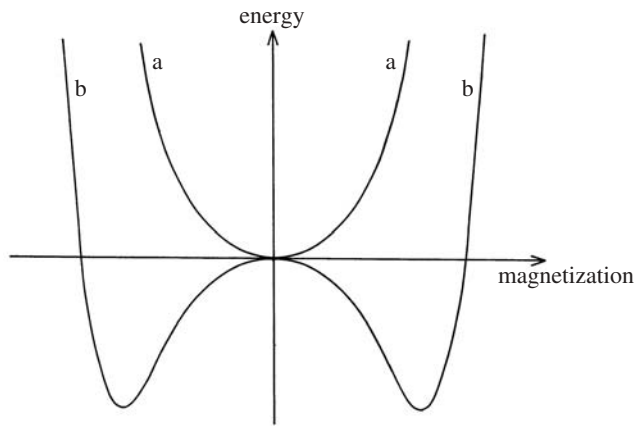


Figure 6.3 Magnetization curves of bar magnet. (a) The temperature is above the Curie point 770 °C and the net magnetization is zero at the potential energy minimum. (b) The temperature is below the Curie point 770 °C and the net magnetization is nonvanishing at the symmetric potential-energy minima.

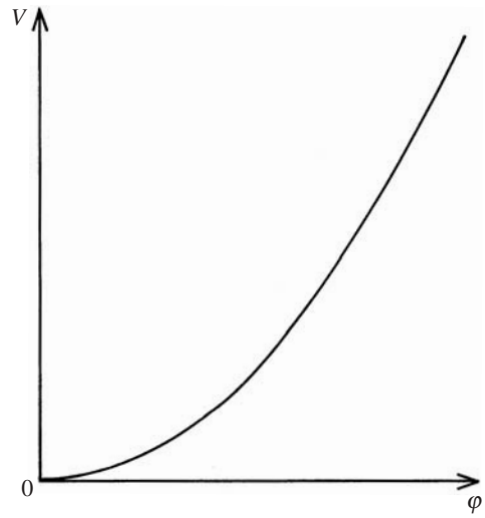


Figure 6.4 Potential energy of the form (6.46) of a real scalar field φ.

the ground state of the bar magnet is in either one of these minima, not in the state of zero magnetization.

The rotational symmetry has then been replaced by the lesser symmetry of parity, or inversion of the magnetization axis. To be exact, this argument actually requires that the bar magnet is infinitely long so that its moment of inertia is infinite. Then no unitary operator can rotate the north pole into the south pole. Note that the potential energy curve in Figure 6.4 has the shape of a polynomial of at least fourth degree.

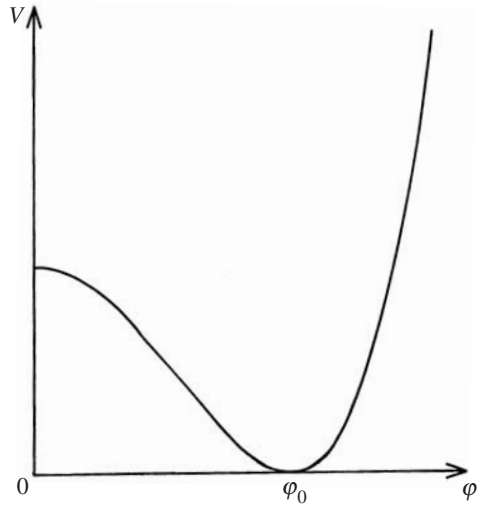


Figure 6.5 Potential energy of the form (6.49) of a real scalar field φ.

Free Scalar Bosons. As a third example of a spontaneously broken symmetry, consider the vacuum filled with a real scalar field $\varphi(x)$, where x stands for the space-time coordinates. Recall that the electric field is a vector—it has a direction. A scalar field is like temperature: it may vary as a function of x, but it has no direction.

If the potential energy in the vacuum is described by the parabolic curve in Figure 6.4, the total energy may be written

$$\tfrac{1}{2}(\nabla\varphi)^2 + \tfrac{1}{2}m^2\varphi^2. \tag{6.46}$$

Here the first term is the kinetic energy, which does not interest us. The second term is the potential energy $V(\varphi)$, which has a minimum at $\varphi = 0$ if φ is a classical field and m^2 is a positive number. The origin of the $\frac{1}{2}$ factors is of no importance to us.

If φ is a quantum field, it oscillates around the classical ground state $\varphi = 0$ as one moves along some trajectory in space-time. The quantum mechanical ground state is called the *vacuum expectation value* of the field. In this case it is

$$\langle\varphi\rangle = 0. \tag{6.47}$$

One can show that the potential (6.46) corresponds to a freely moving scalar boson of mass m (there may indeed exist such particles).

Another parametrization of a potential with a single minimum at the origin is the fourth-order polynomial

$$V(\varphi) = \tfrac{1}{2}m^2\varphi^2 + \tfrac{1}{4}\lambda\varphi^4, \tag{6.48}$$

where λ is some positive constant.

Physical Scalar Bosons. Let us now study the potential in Figure 6.5, which resembles the curve in Figure 6.3 at temperatures below 770 °C. This clearly

requires a polynomial of at least fourth degree. Let us use a form similar to Equation (6.48),

$$V(\varphi) = -\tfrac{1}{2}\mu^2\varphi^2 + \tfrac{1}{4}\lambda\varphi^4. \tag{6.49}$$

The two minima of this potential are at the field values

$$\varphi_0 = \pm\mu/\sqrt{\lambda}. \tag{6.50}$$

Suppose that we are moving along a space-time trajectory from a region where the potential is given by Equation (6.48) to a region where it is given by Equation (6.49). As the potential changes, the original vacuum (6.47) is replaced by the vacuum expectation value $\langle\varphi_0\rangle$. Regardless of the value of φ at the beginning of the trajectory, it will end up oscillating around φ_0 after a time of the order of μ^{-1}. We say that the original symmetry around the unstable *false vacuum* point at $\varphi = 0$ has been broken spontaneously.

Comparing the potentials (6.49) and (6.46), we see that a physical interpretation of (6.49) would correspond to a free scalar boson with negative squared mass, $-\mu^2$! How can this be physical?

Let us replace φ by $\varphi + \varphi_0$ in the expression (6.49). Then

$$V(\varphi) = -\tfrac{1}{2}\mu^2(\varphi + \varphi_0)^2 + \tfrac{1}{4}\lambda(\varphi + \varphi_0)^4 = \tfrac{1}{2}(3\lambda\varphi_0^2 - \mu^2)\varphi^2 + \cdots. \tag{6.51}$$

The dots indicate that terms of third and fourth order in φ have been left out. Comparing the coefficients of φ^2 in Equations (6.46) and (6.51) and substituting (6.50) for φ_0, we obtain

$$m^2 = 3\lambda\varphi_0^2 - \mu^2 = 3\lambda(\mu/\sqrt{\lambda})^2 - \mu^2 = 2\mu^2. \tag{6.52}$$

Thus we see that the *effective mass* of the field is indeed positive, so it can be interpreted as a physical scalar boson.

The bosons in this model can be considered to move in a vacuum filled with the classical background field φ_0. In a way, the vacuum has been redefined, although it is just as empty as before. The only thing that has happened is that *in the process of spontaneous symmetry breaking, the mass of the scalar boson has changed.* The symmetry breaking has this effect on all particles with which the scalar field interacts, fermions and vector bosons alike.

Electroweak Symmetry Breaking. We shall now apply this model to the case of $\mathrm{SU}(2)_\mathrm{w} \otimes U(1)$ symmetry breaking. There are additional complications because of the group structure, but the principle is the same. The model is really the relativistic generalization of the theory of superconductivity of Ginzburg and Landau.

Let us start on a trajectory in a region of space-time where $\mathrm{SU}(2)_\mathrm{w} \otimes U(1)$ is an exact symmetry. The theory requires four vector bosons as we have seen before, but under the exact symmetry they are massless and called B^0, W^+, W^0, W^-. We now invent such a scalar field φ that, as we go along the trajectory into a region where the symmetry is spontaneously broken, the vector bosons obtain their physical values.

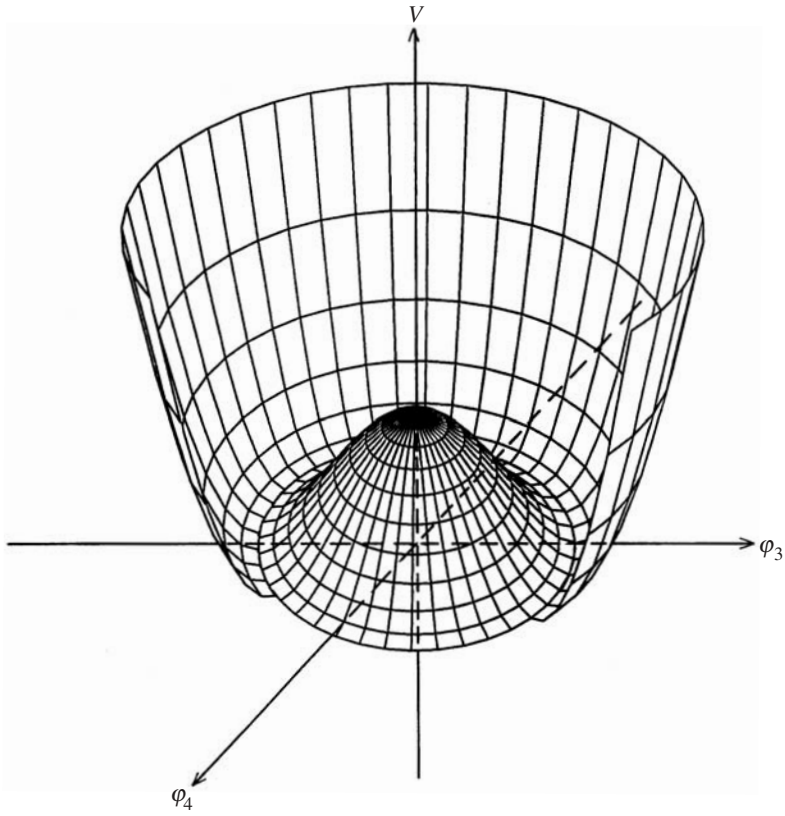


Figure 6.6 'Mexican hat' potential of a complex scalar field φ. All the true vacuum states are located on the minimum Equation (6.54) forming a circle at the bottom of the hat. The false vacuum is on the top of the hat at the centre.

To do this we use a trick invented by *Peter Higgs*. We choose the *Higgs field* $\boldsymbol{\varphi}$ to be a complex scalar $\mathrm{SU}(2)_\mathrm{w}$-doublet,

$$\varphi = \begin{pmatrix} \varphi_1 + \mathrm{i}\varphi_2 \\ \varphi_3 + \mathrm{i}\varphi_4 \end{pmatrix}. \tag{6.53}$$

The vector bosons interact with the four real components φ_i of the $\mathrm{SU}(2)_\mathrm{w}$-symmetric field φ. The false vacuum corresponds to the state $\boldsymbol{\varphi} = 0$, or

$$\varphi_1 = \varphi_2 = \varphi_3 = \varphi_4 = 0.$$

The true vacuum, which has a lower potential energy than the false vacuum, corresponds to the state

$$\varphi_1 = \varphi_2 = 0, \qquad \varphi_3^2 + \varphi_4^2 = \text{const.} > 0. \tag{6.54}$$

This potential is like the one in Figure 6.5, but rotated around the V-axis. Thus it has rotational symmetry like a Mexican hat (see Figure 6.6). All values of the potential on the circle at the bottom are equally good.

If on our trajectory through space-time we come to a point where the field has a special value such as $\varphi_4 = 0$, $\varphi_3 > 0$, then the rotational symmetry of the Mexican hat is spontaneously broken. Just as in the case of the potential (6.51), the scalar field becomes massive, corresponding to one freely moving *Higgs boson*, H^0, in a redefined vacuum. As a consequence, the vector bosons W^+, W^- interacting with the scalar field also become massive (80 GeV). The two neutral fields B^0, W^0 form the linear combinations

$$\gamma = B^0 \cos\theta_w + W^0 \sin\theta_w, \tag{6.55}$$

$$Z^0 = -B^0 \sin\theta_w + W^0 \cos\theta_w, \tag{6.56}$$

of which Z^0 becomes massive (91 GeV), whereas our ordinary photon γ remains massless. γ remains massless is because it is electroweak-neutral (T_3-neutral), so it does not feel the electroweak Higgs field.

Thus the Higgs boson explains the spontaneous breaking of the $SU(2)_w \otimes U(1)$ symmetry. The H^0 mass is

$$m_\varphi = 2\sqrt{\lambda} \times 246 \text{ GeV}. \tag{6.57}$$

Unfortunately, the value of λ is unknown, so this very precise relation is useless! At the time of writing, the Higgs boson has not yet been found, only a lower limit of 114 GeV can be quoted. But since the standard model works very well, physicists are confident of finding it in the next generation of particle accelerators, if not before.

Since the electroweak symmetry $SU(2)_w \otimes U(1)_{B-L}$ is the direct product of two subgroups, $U(1)_{B-L}$ and $SU(2)_w$, it depends on two coupling constants g_1, g_2 associated with the two factor groups. Their values are not determined by the symmetry, so they could in principle be quite different. This is a limitation to the electroweak symmetry. A more symmetric theory would depend on just one coupling constant g. This is one motivation for the search for a GUT which would encompass the electroweak symmetry and the colour symmetry, and which would be more general than their product (6.41).

At the point of spontaneous symmetry breaking several parameters of the theory obtain specific values. It is not quite clear where the different masses of all the quarks and leptons come from, but symmetry breaking certainly plays a role. Another parameter is the so-called *Weinberg angle* θ_w, which is related to the coupling constant of the $SU(2)_w$ subgroup. Its value is not fixed by the electroweak theory, but it is expected to be determined in whatever GUT may be valid.

6.6 Primeval Phase Transitions and Symmetries

The primeval Universe may have developed through phases when some symmetry was exact, followed by other phases when that symmetry was broken. The early cosmology would then be described by a sequence of *phase transitions*. Symmetry breaking may occur through a *first-order phase transition*, in which the field tunnels through a potential barrier, or through a *second-order phase transition*, in

which the field evolves smoothly from one state to another, following the curve of the potential.

Temperature. An important bookkeeping parameter at all times is the temperature, T. When we follow the history of the Universe as a function of T, we are following a trajectory in space-time which may be passing through regions of different vacua. In the simple model of symmetry breaking by a real scalar field, φ, having the potential (6.49), the T-dependence may be put in explicitly, as well as other dependencies (denoted by 'etc.'),

$$V(\varphi, T, \text{etc.}) = -\tfrac{1}{2}\mu^2\varphi^2 + \tfrac{1}{4}\lambda\varphi^4 + \tfrac{1}{8}\lambda T^2\varphi^2 + \cdots . \tag{6.58}$$

As time decreases T increases, the vacuum expectation value ϕ_0 decreases, so that finally, in the early Universe, the true minimum of the potential is the trivial one at $\varphi = 0$. This occurs above a *critical temperature* of

$$T_c = 2\mu/\sqrt{\lambda}. \tag{6.59}$$

An example of this behaviour is illustrated by the potentials in Figure 6.7. The different curves correspond to different temperatures. At the highest temperature, the only minimum is at $\varphi = 0$ but, as the temperature decreases, a new minimum develops spontaneously. If there is more than one minimum, only one of these is stable. A classical example of an unstable minimum is a steam bubble in boiling water.

Early History. Let us now construct a possible scenario for the early history of the Universe. We start at the same point as we did in Section 5.3 (about 1 μs after the Big Bang, when the energy of the particles in thermal equilibrium was about 200 MeV), but now we let time run backwards. For the different scales we can refer to Figure 5.9.

***E* ≈ 200 MeV.** At this temperature, the phase transition between low-energy hadronic physics and QCD occurs: the individual nucleons start to overlap and 'melt' into asymptotically free quarks, forming *quark matter*. In this dense medium the separation between quarks in the nucleons decreases, and the interaction between any two quarks in a nucleon is screened by their interaction with quarks in neighbouring nucleons. Quark matter may still exist today in the core of cold but dense stellar objects such as neutron stars.

The colour symmetry $SU(3)_c$ is valid at this temperature, but it is in no way apparent, because the hadrons are colour-neutral singlets. The colour force mediated by gluons is also not apparent: a vestige of QCD from earlier epochs remains in the form of strong interactions between hadrons. It appears that this force is mediated by mesons, themselves quark bound states.

Above 200 MeV, the particles contributing to the effective degrees of freedom g_*, introduced in Section 5.4, are the photon, three charged leptons, three neutrinos (not counting their three inert right-handed components), the six quarks with

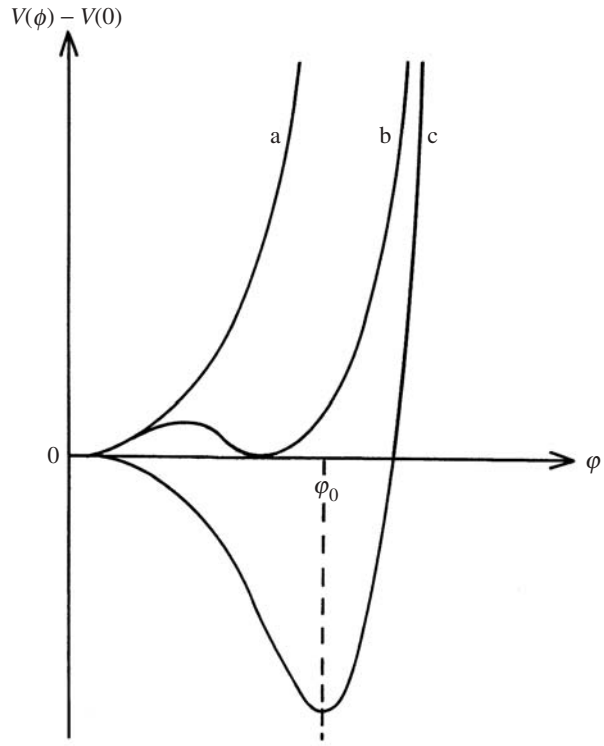


Figure 6.7 Effective scalar potentials. The different curves correspond to different temperatures. At the highest temperature (a) the only minimum is at $\varphi = 0$ but, as the temperature decreases (b), a new minimum develops spontaneously. Finally, in (c), a stable minimum is reached at φ_0.

three colours each, the gluon of eight colours and two spin states, the scalar Higgs boson H^0, and the vector gauge bosons $W^{\pm}$, Z^0. Thus Equation (5.52) is replaced by

$$g_* = 2 + 3 \times \tfrac{7}{2} + 3 \times \tfrac{7}{4} + 6 \times 3 \times \tfrac{7}{2} + 8 \times 2 + 1 + 3 \times 3 = 106.75 \qquad (6.60)$$

This large value explains the (arbitrarily steep) drop at the QCD–hadron phase transition at 200 MeV in Figure 5.6 [5].

There is no trace of the weak-isospin symmetry $SU(2)_W$, so the weak and electromagnetic interactions look quite different. Their strengths are very different, and the masses of the leptons are very different. Only the electromagnetic gauge symmetry $U(1)$ is exactly valid, as is testified to by the conservation of electric charge.

1 GeV $\lesssim$ *E* $\lesssim$ 100 GeV. As the temperature increases through this range, the unity of the weak and electromagnetic forces as the electroweak interaction becomes progressively clearer. The particles contributing to the effective degrees

of freedom in the thermal soup are quarks, gluons, leptons and photons. All the fermions are massive.

The electroweak symmetry $\mathrm{SU}(2)_\mathrm{W} \otimes \mathrm{U}(1)_{B-L}$ is broken, as is testified to by the very different quark masses and lepton masses. The electroweak force is mediated by massless photons and virtual $\mathrm{W}^\pm$, Z^0 vector bosons. The latter do not occur as free particles, because the energy is still lower than their rest masses near the upper limit of this range. The $\mathrm{SU}(3)_\mathrm{c} \otimes \mathrm{U}(1)$ symmetry is of course exact, and the interactions of quarks are ruled by the colour force.

***E* ≈ 100 GeV.** This is about the rest mass of the W and Z, so they freeze out of thermal equilibrium. The Higgs boson also freezes out about now, if it has not already done so at higher temperature. Our ignorance here is due to the lack of experimental information about the mass.

There is no difference between weak and electromagnetic interactions: there are charged-current electroweak interactions mediated by the $\mathrm{W}^\pm$, and neutral-current interactions mediated by the Z^0 and γ. However, the electroweak symmetry is imperfect because of the very different masses.

***E* ≈ 1 TeV.** Up to this energy, our model of the Universe is fairly reliable, because this is the limit of present-day laboratory experimentation. Here we encounter the phase transition between exact and spontaneously broken $\mathrm{SU}(2)_\mathrm{W} \otimes \mathrm{U}(1)_{B-L}$ symmetry. The end of electroweak unification is marked by the massification of the vector boson fields, the scalar Higgs fields and the fermion fields.

One much discussed extension to the standard model is *supersymmetry* (SUSY). This brings in a large number of new particles, some of which should be seen in this temperature range. In this theory there is a conserved multiplicative quantum number, *R parity*, defined by

$$R = (-1)^{3B+L+2s}, \tag{6.61}$$

where B, L and s are baryon number, lepton number and spin, respectively. All known particles have $R = +1$, but the theory allows an equal number of supersymmetric partners, *sparticles*, having $R = -1$. Conservation of R ensures that SUSY sparticles can only be produced pairwise, as sparticle–antisparticle pairs. The lightest sparticles must therefore be stable, just as the lightest particles are.

One motivation for introducing this intermediate scale is the *hierarchy problem*: why is m_P so enormously much larger than m_W? And why is V_Coulomb so much larger than V_Newton? SUSY has so many free parameters that it can 'naturally' explain these problems.

1 TeV $\lesssim$ *E* $\lesssim$ $10^{11\text{–}12}$ TeV. The 'standard' symmetry group G_s in Equation (6.41) is an exact symmetry in this range. As we have seen, laboratory physics has led us to construct the 'standard' theory, which is fairly well understood, although experimental information above 1 TeV is lacking. The big question is

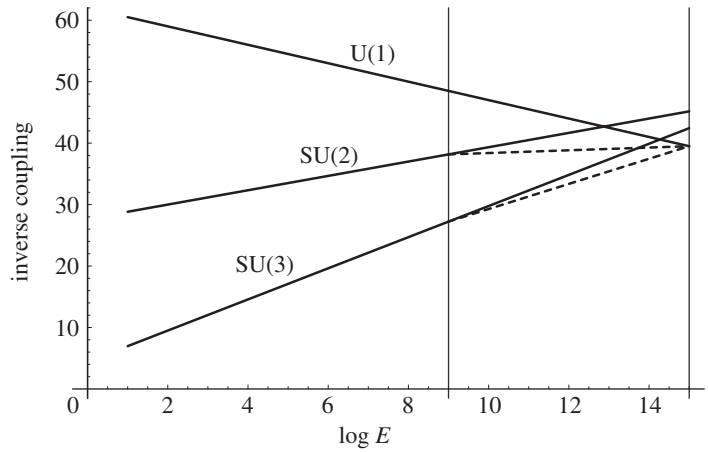


Figure 6.8 The solid lines show the evolution of the inverse of the coupling constants for the symmetry groups $U(1)_{B-L}$, $SU(2)_{\text{electroweak}}$ and $SU(3)_{\text{colour}}$, respectively. The dotted lines illustrate a case when the evolution is broken at an intermediate energy $E_{\text{INT}} = 9$ GeV so that unification occurs at $E_{\text{GUT}} = 15$ GeV.

what new physics will appear in this enormous energy range. The possibility that nothing new appears is called 'the desert'.

The new physics could be a higher symmetry which would be broken at the lower end of this energy range. Somewhere there would then be a phase transition between the exactly symmetric phase and the spontaneously broken phase. Even in the case of a 'desert', one expects a phase transition to GUT at 10^{14} or 10^{15} GeV. One effect which finds its explanation in processes at about 10^{14} GeV is the remarkable absence of antimatter in the Universe.

If the GUT symmetry G_{GUT} breaks down to G_{s} through intermediate steps, the phenomenology could be very rich. For instance, there are *subconstituent models* building leptons and quarks out of elementary particles of one level deeper elementarity. These subconstituents would freeze out at some intermediate energy, condensing into lepton and quark bound states. The forces binding them are in some models called *technicolour forces*.

10^{12} GeV $\lesssim E \lesssim$ 10^{16} TeV. Let us devote some time to GUTs which might be exact in this range. The unification of forces is not achieved very satisfactorily within the G_{s} symmetry. It still is the direct product of three groups, thus there are in principle three independent coupling constants g_1, g_2 and g_3 associated with it. Full unification of the electroweak and colour forces would imply a symmetry group relating those coupling constants to only one g. The specific values of the coupling constants are determined (by accident?) at the moment of spontaneous G_{GUT} breaking.

Below the energy of electroweak unification we have seen that the electromagnetic and weak coupling strengths are quite different. As the energy increases, their relative strengths change. Thus the coupling constants are functions of energy; one says that they are *running*. If one extrapolates the running coupling

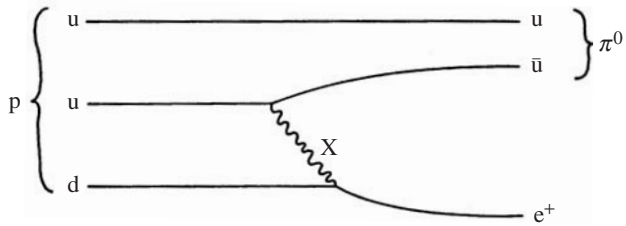


Figure 6.9 Proton decay Feynman diagram.

constants from their known low-energy regime, they almost intersect in one point (see Figure 6.8). The energy of that point is between 10^{13} and 10^{15} GeV, so, if there is a GUT, that must be its unification scale and the scale of the masses of its vector bosons and Higgs bosons.

The fact that the coupling strengths do not run together to exactly one point is actually quite an interesting piece of information. It could imply that there exist intermediate symmetries in the 'desert' between G_{GUT} and G_{s}. Their effect on the lines in Figure 6.8 would be to change their slopes at the intermediate energy. The present extrapolation to GUT energy would then be wrong, and the lines could after all meet exactly at one point. In Figure 6.8 we have illustrated this, choosing $E_{\mathrm{GUT}} = 15$ GeV, and the intermediate energy at $E_{\mathrm{INT}} = 9$ GeV.

As I have pointed out several times, it is not understood why the leptons and the quarks come in three families. The symmetry group requires only one family, but if nature provides us with more, why are there precisely three? This is one incentive to search for larger unifying symmetries. However, *family unification theories* (FUTs) need not be the same as GUT.

The standard model leaves a number of other questions open which one would very much like to have answered within a GUT. There are too many free parameters in G_{s}. Why are the electric charges such as they are? How many Higgs scalars are there? What is the reason for CP violation? The only hint is that CP violation seems to require (but not explain) at least three families of quarks.

Why is parity maximally violated? Could it be that the left–right asymmetry is only a low-energy artefact that disappears at higher energies? There are left–right symmetric models containing G_{s}, in which the right-handed particles interact with their own right-handed vector bosons, which are one or two orders of magnitude heavier than their left-handed partners. Such models then have intermediate unification scales between the G_{s} and G_{GUT} scales.

In a GUT, all the leptons and quarks should be components of the same field. In consequence there must exist *leptoquark* vector bosons, X, which can transform a quark into a lepton in the same way as the colour (i)–anticolour (j) gluons can transform the colour of quarks. This has important consequences for the stability of matter: the quarks in a proton could decay into leptons, for instance, as depicted in Figure 6.9, thereby making protons unstable.

Experimentally we know that the mean life of protons must exceed the age of the Universe by many orders of magnitude. Sensitive underground detectors have been waiting for years to see a single proton decay in large volumes of water.

There are about 10^{33} protons in a detector of 2000 t of water, so if it could see one decay in a year, the proton mean life would be

$$\tau_{\rm p} \gtrsim 10^{33} \text{ yr}. \tag{6.62}$$

This is about the present experimental limit. It sets stringent limits on the possible GUT candidates, and it has already excluded a GUT based on the symmetry group SU(5), which offered the simplest scheme of quark–lepton cohabitation in the same multiplets.

Although all GUTs are designed to answer some of the questions and relate some of the parameters in the standard model, they still have the drawback of introducing large numbers of new particles, vector bosons and Higgs scalars, all of which are yet to be discovered.

E $\lesssim$ **10^{19} GeV.** We have now reached the energy scale where gravitational and quantum effects are of equal importance, so we can no longer do particle physics neglecting gravitation. In quantum mechanics it is always possible to associate the mass of a particle M with a wave having the *Compton wavelength*

$$\lambda = \frac{\hbar}{Mc}. \tag{6.63}$$

In other words, for a particle of mass M, quantum effects become important at distances of the order of λ. On the other hand, gravitational effects are important at distances of the order of the Schwarzschild radius. Equating the two distances, we find the scale at which quantum effects and gravitational effects are of equal importance. This defines the *Planck mass*

$$M_{\rm P} = \sqrt{\hbar c / G} = 1.221 \times 10^{19} \text{ GeV } c^{-2}. \tag{6.64}$$

From this we can derive the Planck energy $M_{\rm P}c^2$ and the *Planck time*

$$t_{\rm P} = \lambda_{\rm P}/c = 5.31 \times 10^{-44} \text{ s}. \tag{6.65}$$

Later on we shall make frequent use of quantities at the Planck scale. The reason for associating these scales with Planck's name is that he was the first to notice that the combination of fundamental constants

$$\lambda_{\rm P} = \sqrt{\hbar G / c^3} = 1.62 \times 10^{-35} \text{ m} \tag{6.66}$$

yielded a natural length scale.

Unfortunately, there is as yet no theory including quantum mechanics and gravitation. Thus we are forced to stop here at the Planck time, a 'mere' 10^{-43} s after the Big Bang, because of a lack of theoretical tools. But we shall come back to these earliest times in connection with models of cosmic inflation in the next chapter and in the final chapter.

6.7 Baryosynthesis and Antimatter Generation

In Section 5.6 we noted that the ratio η of the baryon number density N_{B} to the photon number density N_γ is very small. We can anticipate a value deduced in Equation (8.45)

$$\eta = 6.1 \pm 0.7 \times 10^{-10} \tag{6.67}$$

from deuterium synthesis, CMB and large-scale structures.

The Ratio of Baryons to Photons. Before the baryons were formed (at about 200 MeV), the conserved baryon number B was carried by quarks. Thus the total value of B carried by all protons and neutrons today should equal the total value carried by all quarks. It is very surprising then that η should be so small today, because when the quarks, leptons and photons were in thermal equilibrium there should have been equal numbers of quarks and anti-quarks, leptons and anti-leptons, and the value of B was equal to the total leptonic number L,

$$N_{\mathrm{B}} = N_{\bar{\mathrm{B}}} = N_{\mathrm{L}} \approx N_\gamma \tag{6.68}$$

because of the $\mathrm{U}(1)_{B-L}$ symmetry.

When the baryons and anti-baryons became nonrelativistic, the numbers of baryons and anti-baryons were reduced by annihilation, so N_{B} decreased rapidly by the exponential factor in the Maxwell–Boltzmann distribution (5.43). The number density of photons N_γ is given by Equation (5.5) at all temperatures. Thus the temperature dependence of η should be

$$\eta = \frac{N_{\mathrm{B}}}{N_\gamma} = \frac{\sqrt{2\pi}}{4.808}\left(\frac{m_N}{kT}\right)^{3/2} \mathrm{e}^{-m_N/kT}. \tag{6.69}$$

When the annihilation rate became slower than the expansion rate, the value of η was frozen, and thus comparable to its value today (6.67). The freeze-out occurs at about 20 MeV, when η has reached the value

$$\eta \simeq 6.8 \times 10^{-19}. \tag{6.70}$$

But this is a factor 9×10^8 too small! Thus something must be seriously wrong with our initial condition (6.68).

Baryon–Anti-baryon Asymmetry. The other surprising thing is that no anti-baryons seem to have survived. At temperatures above 200 MeV, quarks and anti-quarks were in thermal equilibrium with the photons because of reactions such as (5.32)–(5.35), as well as

$$\gamma + \gamma \longleftrightarrow \mathrm{q} + \bar{\mathrm{q}}. \tag{6.71}$$

These reactions conserve baryon number, so every quark produced or annihilated is accompanied by one anti-quark produced or annihilated. Since all quarks and anti-quarks did not have time to annihilate one another, it would be reasonable to expect equal number densities of baryons or anti-baryons today.

But we know that the Earth is only matter and not antimatter. The solar wind does not produce annihilations with the Earth or with the other planets, so we know that the Solar System is matter. Since no gamma rays are produced in the interaction of the solar wind with the local interstellar medium, we know that the interstellar medium, and hence our Galaxy, is matter. The main evidence that other galaxies are not composed of antimatter comes from cosmic rays. Our Galaxy contains free protons (cosmic rays) with a known velocity spectrum very near the speed of light. A fraction of these particles have sufficiently high velocities to escape from the Galaxy. These protons would annihilate with antimatter cosmic rays in the intergalactic medium or in collisions with antimatter galaxies if they existed, and would produce characteristic gamma-rays many orders of magnitude more frequently than have been seen. The observed ratio is

$$\frac{N_{\bar{\mathrm{B}}}}{N_{\mathrm{B}}} = 10^{-5}\text{–}10^{-4},$$

depending on the kinetic energy. This small number is fully consistent with all the observed anti-protons having been produced by energetic cosmic rays in the Earth's atmosphere, and it essentially rules out the possibility that other galaxies emit cosmic rays composed of antimatter. There are many other pieces of evidence against antimatter, but the above arguments are the strongest.

We are then faced with two big questions. What caused the large value of η? And why does the Universe not contain antimatter, anti-quarks and positrons. The only reasonable conclusion is that $N_{\bar{\mathrm{B}}}$ and N_{B} must have started out slightly different while they were in thermal equilibrium, by the amount

$$N_{\mathrm{B}} - N_{\bar{\mathrm{B}}} \simeq \eta N_\gamma. \tag{6.72}$$

Subsequently most anti-baryons were annihilated, and the small excess ηN_γ of baryons is what remained. This idea is fine, but the basic problem has not been removed; we have only pushed it further back to earlier times, to some early B$\bar{\mathrm{B}}$-asymmetric phase transition.

Primeval Asymmetry Generation. Let us consider theories in which a B$\bar{\mathrm{B}}$-asymmetry could arise. For this three conditions must be met.

First, the theory must contain reactions violating baryon number conservation. Grand unified theories are obvious candidates for a reason we have already met in Section 6.6. We noted there that GUTs are symmetric with respect to leptons and quarks, because they are components of the same field and GUT forces do not see any difference. Consequently, GUTs contain leptoquarks X, Y which transform quarks into leptons. Reactions involving X, Y do explicitly violate both baryon number conservation and lepton number conservation since the quarks have $B = \frac{1}{3}$, $L_i = 0$, whereas leptons have $B = 0$, $L_i = 1$, where $i = \mathrm{e}, \mu, \tau$. The baryon and lepton numbers then change, as for instance in the decay reactions

$$X \longrightarrow \mathrm{e}^- + \mathrm{d}, \qquad \Delta B = +\tfrac{1}{3}, \qquad \Delta L_{\mathrm{e}} = 1, \tag{6.73}$$

$$X \longrightarrow \bar{\mathrm{u}} + \bar{\mathrm{u}}, \qquad \Delta B = -\tfrac{2}{3}. \tag{6.74}$$

Secondly, there must be C and CP violation in the theory, as these operators change baryons into anti-baryons and leptons into anti-leptons. If the theory were C and CP symmetric, even the baryon-violating reactions (6.73) and (6.74) would be matched by equally frequently occurring reactions with opposite ΔB, so no net B$\bar{\text{B}}$-asymmetry would result. In fact, we want baryon production to be slightly more frequent than anti-baryon production.

Thirdly, we must require these processes to occur out of thermal equilibrium. In thermal equilibrium there is no net production of baryon number, because the reactions (6.73) and (6.74) go as frequently in the opposite direction. Hence the propitious moment is the phase transition when the X-bosons are freezing out of thermal equilibrium and decay. If we consult the timetable in Section 6.6, this would happen at about 10^{14} GeV: the moment for the phase transition from the GUT symmetry to its spontaneously broken remainder.

The GUT symmetry offers a good example, which we shall make use of in this Section, but it is by no means obvious that GUT is the symmetry we need and that the phase transition takes place at GUT temperature. It is more likely that we have the breaking of a symmetry at a lower energy, such as supersymmetry.

Leptoquark Thermodynamics. Assuming the GUT symmetry, the scenario is therefore the following. At some energy $E_{\text{X}} = kT_{\text{X}}$ which is of the order of the rest masses of the leptoquark bosons X,

$$E_{\text{X}} \simeq M_{\text{X}}c^2, \tag{6.75}$$

all the X, Y vector bosons, the Higgs bosons, the W, B vector bosons of Equations (6.55) and (6.56), and the gluons are in thermal equilibrium with the leptons and quarks. The number density of each particle species is about the same as the photon number density, and the relations (6.68) hold.

When the age of the Universe is still young, as measured in Hubble time τ_{H}, compared with the mean life $\tau_{\text{X}} = \Gamma_{\text{X}}^{-1}$ of the X bosons, there are no X decays and therefore no net baryon production. The X bosons start to decay when

$$\Gamma_{\text{X}} \lesssim \tau_{\text{H}}^{-1} = H. \tag{6.76}$$

This is just like the condition (5.63) for the decoupling of neutrinos. The decay rate Γ_{X} is proportional to the mass M_{X},

$$\Gamma_{\text{X}} = \alpha M_{\text{X}}, \tag{6.77}$$

where α is essentially the coupling strength of the GUT interaction. It depends on the details of the GUT and the properties of the X boson.

We next take the temperature dependence of the expansion rate H from Equations (5.49) and (5.51). Replacing the Newtonian constant G by its expression in terms of the Planck mass M_{P}, as given in Equation (6.64), we find

$$H = \sqrt{\frac{4\pi\hbar a}{3c}g_*(T)}\frac{T^2}{M_{\text{P}}}. \tag{6.78}$$

Substituting this H and the expression (6.77) into the condition (6.76), the Universe is out of equilibrium when

$$AM_{\mathrm{X}} \lesssim \sqrt{g_*(T)}\,\frac{T^2}{M_{\mathrm{P}}} \quad \text{at } T = M_{\mathrm{X}}, \tag{6.79}$$

where all the constants have been lumped into A. Solving for the temperature squared, we find

$$T^2 \gtrsim \frac{AM_{\mathrm{X}}M_{\mathrm{P}}}{\sqrt{g_*(T)}}. \tag{6.80}$$

At temperature T_{X}, the effective degrees of freedom g_* are approximately 100. The condition (6.80) then gives a lower limit to the X boson mass,

$$M_{\mathrm{X}} \gtrsim A'\frac{M_{\mathrm{P}}}{\sqrt{g_*(T)}} = A'\frac{1.2\times 10^{19}\ \mathrm{GeV}}{\sqrt{10\,675}} \simeq A'\times 10^{18}\ \mathrm{GeV}, \tag{6.81}$$

where A' includes all constants not cited explicitly.

Thus, if the mass M_{X} is heavier than $A'\times 10^{18}$ GeV, the X bosons are stable at energies above M_{X}. Let us assume that this is the case. As the energy drops below M_{X}, the X and $\bar{\mathrm{X}}$ bosons start to decay, producing the net baryon number required. The interactions must be such that the decays really take place out of equilibrium, that is, the temperature of decoupling should be above M_{X}. Typically, bosons decouple from annihilation at about $M_{\mathrm{X}}/20$, so it is not trivial to satisfy this requirement.

Let us now see how C and CP violation can be invoked to produce a net B$\bar{\mathrm{B}}$-asymmetry in X and $\bar{\mathrm{X}}$ decays. We can limit ourselves to the case when the only decay channels are (6.73) and (6.74), and correspondingly for the $\bar{\mathrm{X}}$ channels. For these channels we tabulate in Table A.7 the net baryon number change ΔB and the ith branching fractions $\Gamma(X \to \text{channel } i)/\Gamma(X \to \text{ all channels})$ in terms of two unknown parameters r and $\bar{r}$.

The baryon number produced in the decay of one pair of X, $\bar{\mathrm{X}}$ vector bosons weighted by the different branching fractions is then

$$\Delta B = r\Delta B_1 + (1-r)\Delta B_2 + \bar{r}\Delta B_3 + (1-\bar{r})\Delta B_4 = \bar{r} - r. \tag{6.82}$$

If C and CP symmetry are violated, r and $\bar{r}$ are different, and we obtain the desired result $\Delta B \neq 0$. Similar arguments can be made for the production of a net lepton–anti-lepton asymmetry, but nothing is yet known about leptonic CP violation.

Suppose that the number density of X and $\bar{\mathrm{X}}$ bosons is N_{X}. We now want to generate a net baryon number density

$$N_{\mathrm{B}} = \Delta B N_{\mathrm{X}} \simeq \Delta B N_\gamma$$

by the time the Universe has cooled through the phase transition at T_{GUT}. After that the baryon number is absolutely conserved and further decrease in N_{B} only follows the expansion. However, the photons are also bosons, so their absolute number is not conserved and the value of η may be changing somewhat. Thus, if

we want to confront the baryon production ΔB required at $T_{\rm GUT}$ with the present-day value of η, a more useful quantity is the baryon number per unit entropy $N_{\rm B}/S$. Recall that the entropy density of photons is

$$s = 1.80 g_*(T) N_\gamma \tag{6.83}$$

from Equation (5.66). At temperature $T_{\rm GUT}$ the effective degrees of freedom were shown in Equation (6.60) to be 106.75 (in the standard model, not counting lepto-quark degrees of freedom), so the baryon number per unit entropy is

$$\frac{N_{\rm B}}{S} = \frac{\Delta B}{1.80 g_*(T_{\rm GUT})} \simeq \frac{\Delta B}{180}. \tag{6.84}$$

Clearly this ratio scales with $g_*^{-1}(T)$. Thus, to observe a present-day value of η at about the value in (6.67), the GUT should be chosen such that it yields

$$\Delta B = \frac{g_*(T_{\rm GUT})}{g_*(T_0)} \eta \simeq \frac{106.75}{3.36} \eta \approx 1.9 \times 10^{-8}, \tag{6.85}$$

making use of the $g_*(T)$ values in Equations (5.73) and (6.60). This is within the possibilities of various GUTs.

One may of course object that this solution of the *baryosynthesis* problem is only speculative, since it rests on the assumption that nature exhibits a suitable symmetry. At the beginning of this section, we warned that the GUT symmetry did not necessarily offer the best phase transition mechanism for baryosynthesis. The three conditions referred to could perhaps be met at some later phase transition. The reason why the GUT fails is to be found in the scenarios of cosmic inflation (Chapter 7). The baryon asymmetry produced at $T_{\rm GUT}$ is subsequently washed out when the Universe reheats to $T_{\rm GUT}$ at the end of inflation.

The search for another mechanism has turned to the electroweak phase transition at about 100 GeV. The 'minimal standard model' of electroweak interactions cannot generate an asymmetry but, if the correct electroweak theory could be more general. New possibilities arise if all three neutrino species oscillate and violate CP, or if one turns to the 'minimal supersymmetric standard model'. At the expanding interface of the broken symmetry phase, the baryon–anti-baryon asymmetry could be generated via complex CP-violating reflections, transmissions and interference phenomena between fermionic excitations. Thus the existence of baryons is an indication that physics indeed has to go beyond the 'minimal standard model'.

Problems

1. Are the raising and lowering operators S_+ and S_- unitary or Hermitian?
2. Verify the commutation relations (6.19) for the Pauli spin matrices (6.18).
3. Write down the weak hypercharge operator Y (6.32) in matrix form.

4. The s quark is assigned the value of strangeness $S = -1$. The relation (6.25) for all nonstrange hadrons reads $Q = \frac{1}{2}B + I_3$. Generalize the relation (6.25) to include strangeness so that it holds true for the K mesons defined in Equations (6.34). Note that, of all the quark–anti-quark systems possible with three quarks, only five are listed in Equations (6.34). Write down the quark content of the remaining systems.
5. Show by referring to the quark structure that the K mesons are not eigenstates of the C operator.
6. All known baryons are qqq-systems. Use the u, d, s quarks to compose the 27 ground-state baryons, and derive their charge and strangeness properties. Plot these states in $(I_3, B + S)$-space.
7. Is the parity operator P defined in Equation (6.42) Hermitian?
8. One can define the states $\mathrm{K_L} = \mathrm{K}^0 - \bar{\mathrm{K}}^0$ and $\mathrm{K_S} = \mathrm{K}^0 + \bar{\mathrm{K}}^0$. Prove that these are eigenstates of the CP operator. These states decay dominantly to 2π and 3π. Which state decays to which, and why? What does this imply for the relative lifetimes of the K_S and K_L?
9. Derive a value of weak hypercharge $Y = B - L$ for the X boson from reactions (6.72) and (6.73).

Chapter Bibliography

[1] Chaichian, M. and Nelipa, N. F. 1984 *Introduction to gauge field theories*. Springer.
[2] Kolb, E. W. and Turner, M. S. 1990 *The early Universe*. Addison-Wesley, Reading, MA.
[3] Collins, P. D. B., Martin, A. D. and Squires, E. J. 1989 *Particle physics and cosmology*. John Wiley & Sons, New York.
[4] Linde, A. 1990 *Particle physics and inflationary cosmology*. Harwood Academic Publishers, London.
[5] Coleman, T. S. and Roos, M. 2003 *Phys. Rev.* D **68**, 027702.

7

Cosmic Inflation

The standard FLRW Big Bang model describes an adiabatically expanding Universe, having a beginning of space and time with nearly infinite temperature and density. This model has, as so far presented, been essentially a success story. But the Big Bang assumes very problematic initial conditions: for instance, where did the 10^{90} particles which make up the visible Universe come from? We are now going to correct that optimistic picture and present a remedy: cosmic inflation.

In Section 7.1 we shall discuss problems caused by the expansion of space-time: the *horizon problem* related to its size at different epochs, the *monopole problem* associated with possible topological defects, and the *flatness problem* associated with its metric.

In Section 7.2 we shall study a now classical scenario to solve these problems, called *'old' inflation*. In this scenario the Universe traversed an epoch when a scalar field with negative pressure caused a de Sitter-like expansion, and terminated it with a huge entropy increase, in violation of the law of entropy conservation (Equation (5.13)). Although this scenario was qualitatively possible, it had quantitative flaws which were in part alleviated in 'new' inflation.

In Section 7.3 we discuss the scenario of *chaotic inflation*, which introduces a bubble universe in which we inhabit one bubble, totally unaware of other bubbles. The inflationary mechanism is the same In each bubble, but different parameter values may produce totally different universes. Since our bubble must be just right for us to exist in, this model is a version of the *Anthropic principle*. We close this section with a discussion of the predictions of inflation, to be tested in Chapters 8 and 9.

In Section 7.4 we reconnect to the quintessence models of Section 4.3, learning how the primordial inflaton field could be connected to the dark energy field of today.

In Section 7.5 we turn our attention to a speculative alternative to inflation, a cyclic universe containing dark energy as a driving force.

Introduction to Cosmology Third Edition by Matts Roos
© 2003 John Wiley & Sons, Ltd ISBN 0 470 84909 6 (cased) ISBN 0 470 84910 X (pbk)

7.1 Paradoxes of the Expansion

Particle Horizons. Recall the definition of the particle horizon, Equation (2.47), which in a spatially flat metric is

$$\chi_{\rm ph} = \sigma_{\rm ph} = c\int_{t_{\rm min}}^{t_0} \frac{{\rm d}t}{R(t)} = c\int_{R_{\rm min}}^{R_0} \frac{{\rm d}R}{R\dot{R}}. \tag{7.1}$$

This was illustrated in Figure 2.1. In expanding Friedmann models, the particle horizon is finite. Let us go back to the derivation of the time dependence of the scale factor $R(t)$ in Equations (4.39)–(4.41). At very early times, the mass density term in the Friedmann equation (4.4) dominates over the curvature term (we have also called it the vacuum-energy term),

$$\frac{kc^2}{R^2} \ll \frac{8\pi G}{3}\rho. \tag{7.2}$$

This permits us to drop the curvature term and solve for the Hubble parameter,

$$\frac{\dot{R}}{R} = H(t) = \left(\frac{8\pi G}{3}\rho\right)^{1/2}. \tag{7.3}$$

Substituting this relation into Equation (7.1) we obtain

$$\sigma_{\rm ph} = c\int_{R_{\rm min}}^{R_0} \frac{{\rm d}R}{R^2(\dot{R}/R)} = \left(\frac{3c^2}{8\pi G}\right)^{1/2}\int_{R_{\rm min}}^{R_0} \frac{{\rm d}R}{R^2\sqrt{\rho}}. \tag{7.4}$$

In a radiation-dominated Universe, ρ scales like R^{-4}, so the integral on the right converges in the lower limit $R_{\rm min} = 0$, and the result is that the particle horizon is finite:

$$\sigma_{\rm ph} \propto \int_0^{R_0} \frac{{\rm d}R}{R^2R^{-2}} = R_0. \tag{7.5}$$

Similarly, in a matter-dominated Universe, ρ scales like R^{-3}, so the integral also converges, now yielding $\sqrt{R_0}$. Note that an observer living at a time $t_1 < t_0$ would see a smaller particle horizon, $R_1 < R_0$, in a radiation-dominated Universe or $\sqrt{R_1} < \sqrt{R_0}$ in a matter-dominated Universe.

Suppose however, that the curvature term or a cosmological constant dominates the Friedmann equation at some epoch. Then the conditions (4.35) and (4.36) are not fulfilled; on the contrary, we have a negative net pressure

$$p < -\tfrac{1}{3}\rho c^2. \tag{7.6}$$

Substituting this into the law of energy conservation (4.24) we find

$$\dot{\rho} < -2\frac{\dot{R}}{R}\rho. \tag{7.7}$$

This can easily be integrated to give the R dependence of ρ,

$$\rho < R^{-2}. \tag{7.8}$$

Inserting this dependence into the integral on the right-hand side of Equation (7.4) we find

$$\sigma_{\rm ph} \propto \int_{R_{\rm min}}^{R_0} \frac{{\rm d}R}{R^2\sqrt{R^{-2}}} = \int_{R_{\rm min}}^{R_0} \frac{{\rm d}R}{R}, \tag{7.9}$$

an integral which does not converge at the limit $R_{\rm min} = 0$. Thus the particle horizon is not finite in this case. But it is still true that an observer living at a time $t_1 < t_0$ would see a particle horizon that is smaller by $\ln R_0 - \ln R_1$.

Horizon Problem. A consequence of the finite age t_0 of the Universe is that the particle horizon today is finite and larger than at any earlier time t_1. Also, the spatial width of the past light cone has grown in proportion to the longer time perspective. Thus the spatial extent of the Universe is larger than that our past light cone encloses today; with time we will become causally connected with new regions as they move in across our horizon. This renders the question of the full size of the whole Universe meaningless—the only meaningful size being the diameter of its horizon at a given time.

In Chapter 5 we argued that thermal equilibrium could be established throughout the Universe during the radiation era because photons could traverse the whole Universe and interactions could take place in a time much shorter than a Hubble time. However, there is a snag to this argument: the conditions at any space-time point can only be influenced by events within its past light cone, and the size of the past light cone at the time of last scattering ($t_{\rm LSS}$) would appear to be far too small to allow the currently observable Universe to come into thermal equilibrium.

Since the time of last scattering, the particle horizon has grown with the expansion in proportion to the $\frac{2}{3}$-power of time (actually this power law has been valid since the beginning of matter domination at $t_{\rm eq}$, but $t_{\rm LSS}$ and $t_{\rm eq}$ are nearly simultaneous). The net effect is that the particle horizon we see today covers regions which were causally disconnected at earlier times.

At the time of last scattering, the Universe was about 1065 times smaller than it is now ($z_{\rm LSS} \approx 1065$), and the time perspective back to the Big Bang was only the fraction $t_{\rm LSS}/t_0 \simeq 2.3 \times 10^{-5}$ of our perspective. If we assume that the Universe was radiation dominated for all the time prior to $t_{\rm LSS}$, then, from Equation (4.40), $R(t) \propto \sqrt{t}$. The particle horizon at the LSS, $\sigma_{\rm ph}$, is obtained by substituting $R(t) \propto (t_{\rm LSS}/t)^{-1/2}$ into Equation (2.37) and integrating from zero time to $t_{\rm LSS}$:

$$\sigma_{\rm ph} \propto \int_0^{t_{\rm LSS}} {\rm d}t \left(\frac{t_{\rm LSS}}{t}\right)^{1/2} = 2t_{\rm LSS}. \tag{7.10}$$

It is not very critical what we call 'zero' time: the lower limit of the integrand has essentially no effect even if it is chosen as late as $10^{-4}t_{\rm LSS}$.

The event horizon at the time of last scattering, $\sigma_{\rm eh}$, represents the extent of the Universe we can observe today as light from the LSS (cf. Figures 2.1 and 7.1), since we can observe no light from before the LSS. On the other hand, the particle horizon $\sigma_{\rm ph}$ represents the extent of the LSS that could have come into causal

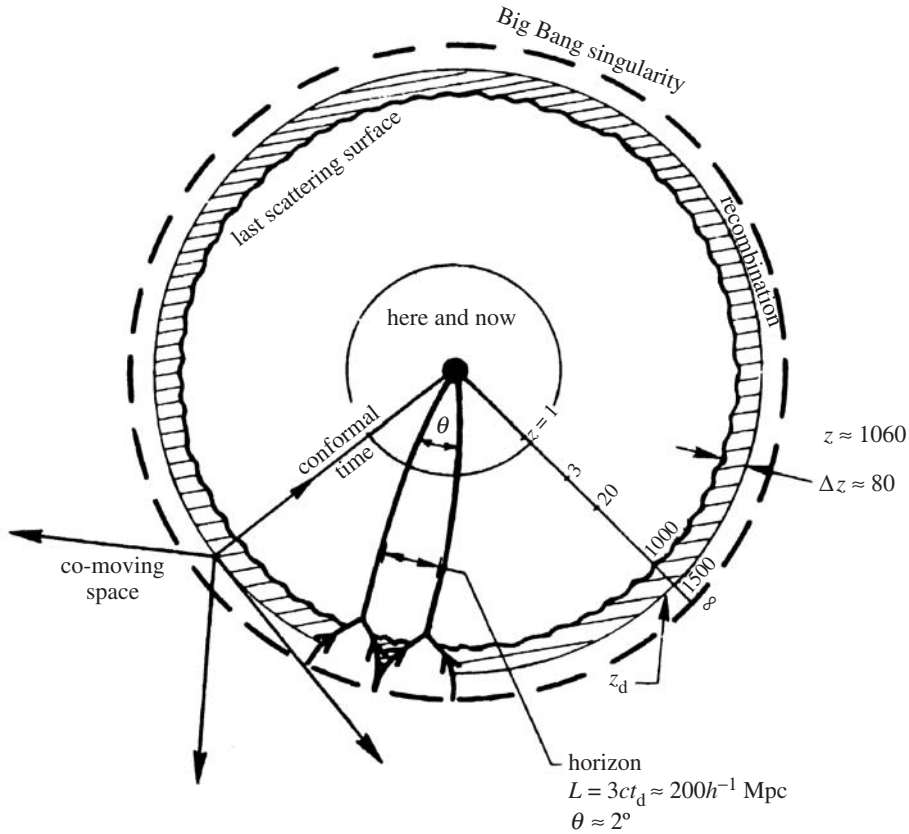


Figure 7.1 A co-moving space/conformal time diagram of the Big Bang. The observer (here and now) is at the centre. The Big Bang singularity has receded to the outermost dashed circle, and the horizon scale is schematically indicated at last scattering. It corresponds to an arc of angle θ today. Reproduced from reference [1] by permission of J. Silk and Macmillan Magazines Ltd.

contact from $t = 0$ to $t_{\rm LSS}$. If the event horizon is larger than the particle horizon, then all the Universe we now see (in particular the relic CMB) could not have been in causal contact by $t_{\rm LSS}$.

The event horizon $\sigma_{\rm eh}$, is obtained by substituting $R(t) \propto (t_{\rm LSS}/t)^{-2/3}$ from Equation (4.39) into Equation (2.49) and integrating over the full epoch of matter domination from $t_{\rm LSS}$ to $t_{\rm max} = t_0$. Assuming flat space, $k = 0$, we have

$$\sigma_{\rm eh} \propto \int_{t_{\rm LSS}}^{t_0} {\rm d}t \left(\frac{t_{\rm LSS}}{t}\right)^{2/3} = 3t_{\rm LSS}\left[\left(\frac{t_0}{t_{\rm LSS}}\right)^{1/3} - 1\right]. \tag{7.11}$$

Let us take $t_{\rm LSS} = 0.35$ Myr and $t_0 = 15$ Gyr. Then the LSS particle horizon $\sigma_{\rm ph}$ is seen today as an arc on the periphery of our particle horizon, subtending an angle

$$\frac{180}{\pi}\left[\frac{\sigma_{\rm ph}}{\sigma_{\rm eh}}\right]_{\rm LSS} \simeq 1.12^\circ. \tag{7.12}$$

This is illustrated in Figure 7.1, which, needless to say, is not drawn to scale. It follows that the temperature of the CMB radiation coming from any 1° arc could not have been causally connected to the temperature on a neighbouring arc, so there is no reason why they should be equal. Yet the Universe is homogeneous and isotropic over the full 360°.

This problem can be avoided, as one sees from Equations (7.6)–(7.9), when the net pressure is negative, for example, when a cosmological constant dominates. In such a case, $R(t) \propto \mathrm{e}^{\mathrm{const.}\cdot t}$ (the case $w = -1$ in Equation (4.38)). If a cosmological constant dominates for a finite period, say between t_1 and $t_2 < t_{\mathrm{LSS}}$, then a term $\mathrm{e}^{\mathrm{const.}\cdot(t_2-t_1)}$ enters into (7.10). This term can be large, allowing a reordering of horizons to give $\sigma_{\mathrm{ph}} > \sigma_{\mathrm{eh}}$.

The age of the Universe at temperature 20 MeV was $t = 2$ ms and the distance scale $2ct$. The amount of matter inside that horizon was only about $10^{-5}M_\odot$, which is very far from what we see today: matter is separated into galaxies of mass $10^{12}M_\odot$. The size of present superclusters is so large that their mass must have been assembled from vast regions of the Universe which were outside the particle horizon at $t = 2$ ms. But then they must have been formed quite recently, in contradiction to the age of the quasars and galaxies they contain. This paradox is the horizon problem.

The lesson of Equations (7.4)–(7.9) is that we can get rid of the horizon problem by choosing physical conditions where the net pressure is negative, either by having a large curvature term or a dominating cosmological term or some large scalar field which acts as an effective cosmological term. We turn to the latter case in Section 7.2.

GUT Phase Transition. Even more serious problems emerge as we approach very early times. At GUT time, the temperature of the cosmic background radiation was $T_{\mathrm{GUT}} \simeq 1.2 \times 10^{28}$ K, or a factor

$$\frac{T_{\mathrm{GUT}}}{T_0} \simeq 4.4 \times 10^{27}$$

greater than today. This is the factor by which the linear scale $R(t)$ has increased since the time t_{GUT}. If we take the present Universe to be of size $2000h^{-1}$ Mpc $= 6 \times 10^{25}$ m, its linear size was only 2 cm at GUT time.

Note, however, that linear size and horizon are two different things. The horizon size depends on the time perspective back to some earlier time. Thus the particle horizon today has increased since t_{GUT} by almost the square of the linear scale factor, or by

$$\frac{t_0}{t_{\mathrm{GUT}}} = \left(\frac{g_*(T_{\mathrm{GUT}})}{g_*(T_0)}\right)^{1/2}\left(\frac{T_{\mathrm{GUT}}}{T_{\mathrm{LSS}}}\right)^2\left(\frac{T_{\mathrm{LSS}}}{T_0}\right)^{3/2} \simeq 2.6 \times 10^{54}. \tag{7.13}$$

At GUT time the particle horizon was only 2×10^{-29} m. It follows that to arrive at the present homogeneous Universe, the homogeneity at GUT time must have extended out to a distance 5×10^{26} times greater than the distance of causal

contact! Why did the GUT phase transition happen simultaneously in a vast number of causally disconnected regions? Concerning even earlier times, one may ask the same question about the Big Bang. Obviously, this paradox poses a serious problem to the standard Big Bang model.

In all regions where the GUT phase transition was completed, several important parameters—such as the coupling constants, the charge of the electron, and the masses of the vector bosons and Higgs bosons—obtained values which would characterize the present Universe. Recall that the coupling constants are functions of energy, as illustrated in Figure 6.8, and the same is true for particle masses. One may wonder why they obtained the same value in all causally disconnected regions.

The Higgs field had to take the same value everywhere, because this is uniquely dictated by what is its ground state. But one might expect that there would be domains where the phase transition was not completed, so that certain remnant symmetries froze in. The Higgs field could then settle to different values, causing some parameter values to be different. The physics in these domains would then be different, and so the domains would have to be separated by *domain walls*, which are *topological defects* of space-time. Such domain walls would contain enormous amounts of energy and, in isolation, they would be indestructible. Intersecting domain walls would produce other types of topological defects such as *loops* or *cosmic strings* wiggling their way through the Universe. No evidence for topological defects has been found, perhaps fortunately for us, but they may still lurk outside our horizon.

Magnetic Monopoles. A particular kind of topological defect is a *magnetic monopole*. Ordinarily we do not expect to be able to separate the north and south poles of a bar magnet into two independent particles. As is well known, cutting a bar magnet into two produces two dipole bar magnets. Maxwell's equations account for this by treating electricity and magnetism differently: there is an electric source term containing the charge e, but there is no magnetic source term. Thus free electric charges exist, but free magnetic charges do not. Stellar bodies may have large magnetic fields, but no electric fields.

Paul A. M. Dirac (1902–1984) suggested in 1931 that the quantization of the electron charge might be the consequence of the existence of at least one free magnetic monopole with magnetic charge

$$g_{\mathrm{M}} = \frac{1}{2}\frac{\hbar c n}{e} \simeq 68.5 e n, \tag{7.14}$$

where e is the charge of the electron and n is an unspecified integer. This would then modify Maxwell's equations, rendering them symmetric with respect to electric and magnetic source terms. Free magnetic monopoles would have drastic consequences, for instance destroying stellar magnetic fields.

Without going into detail about how frequently monopoles might arise during the GUT phase transition, we assume that there could arise one monopole per

10 horizon volumes

$$N_{\rm M}(t_{\rm GUT}) = 0.1 \times (2 \times 10^{-29}\ {\rm m})^{-3},$$

and the linear scale has grown by a factor 4.4×10^{27}. Nothing could have destroyed them except monopole–anti-monopole annihilation, so the monopole density today should be

$$N_{\rm M}(t_0) \simeq 0.1 \times (4.4 \times 0.02\ {\rm m})^{-3} \simeq 150\ {\rm m}^{-3}. \tag{7.15}$$

This is quite a substantial number compared with the proton density which is at most $0.17\ {\rm m}^{-3}$. Monopoles circulating in the Galaxy would take their energy from the galactic magnetic field. Since the field survives, this sets a very low limit to the monopole flux called the *Parker bound*. Experimental searches for monopoles have not yet become sensitive enough to test the Parker bound, but they are certainly in gross conflict with the above value of $N_{\rm M}$; the present experimental upper limit to $N_{\rm M}$ is 25 orders of magnitude smaller than N_y.

Monopoles are expected to be superheavy,

$$m_{\rm M} \gtrsim \frac{m_{\rm X}}{\alpha_{\rm GUT}} \simeq 10^{16}\ {\rm GeV} \simeq 2 \times 10^{-11}\ {\rm kg}. \tag{7.16}$$

Combining this mass with the number densities (5.82) and (7.15) the density parameter of monopoles becomes

$$\Omega_{\rm M} = \frac{N_{\rm M} m_{\rm M}}{\rho_{\rm c}} \simeq 2.8 \times 10^{17}. \tag{7.17}$$

This is in flagrant conflict with the value of the density parameter in Equation (4.79): yet another paradox. Such a universe would be closed and its maximal lifetime would be only a fraction of the age of the present Universe, of the order of

$$t_{\rm max} = \frac{\pi}{2H_0\sqrt{\Omega_{\rm M}}} \simeq 40\ {\rm yr}. \tag{7.18}$$

Monopoles have other curious properties as well. Unlike the leptons and quarks, which appear to be pointlike down to the smallest distances measured (10^{-19} m), the monopoles have an internal structure. All their mass is concentrated within a core of about 10^{-30} m, with the consequence that the temperature in the core is of GUT scale or more. Outside that core there is a layer populated by the X leptoquark vector bosons, and outside that at about 10^{-17} m there is a shell of W and Z bosons.

The monopoles are so heavy that they should accumulate in the centre of stars where they may collide with protons. Some protons may then occasionally penetrate in to the GUT shell and collide with a virtual leptoquark, which transforms a d quark into a lepton according to the reaction

$$\mathrm{d} + \mathrm{X}_{\rm virtual} \longrightarrow \mathrm{e}^+. \tag{7.19}$$

This corresponds to the lower vertex in the drawing of the proton decay diagram (Figure 6.9). Thus monopoles would destroy hadronic matter at a rate much higher than their natural decay rate. This would catalyse a faster disappearance of baryonic matter and yield a different timescale for the Universe.

Flatness Problem. Recall that in a spatially flat Einstein–de Sitter universe the curvature parameter k vanishes and the density parameter is $\Omega = 1$. This is obvious from Equation (4.11), where k and Ω are related by

$$\Omega - 1 = \frac{kc^2}{\dot{R}^2}.$$

The current value of the total density parameter Ω_0 is of order unity. This does not seem remarkable until one considers the extraordinary fine-tuning required: a value of Ω_0 close to, but not exactly, unity today implies that $\Omega_0(t)$ at earlier times must have been close to unity with incredible precision. During the radiation era the energy density ε_r is proportional to R^{-4}. It then follows from Equation (4.4) that

$$\dot{R}^2 \propto R^{-2}. \tag{7.20}$$

At GUT time, the linear scale was some 10^{27} times smaller than today, and since most of this change occurred during the radiation era

$$\Omega - 1 \propto R^2 \simeq 10^{-54}. \tag{7.21}$$

Thus the Universe at that time must have been flat to within 54 decimal places, a totally incredible situation. If this were not so the Universe would either have reached its maximum size within one Planck time (10^{-43} s), and thereafter collapsed into a singularity, or it would have dispersed into a vanishingly small energy density. The only natural values for Ω are therefore 0, 1 or infinity, whereas to generate a universe surviving for several Gyr without a Ω value of exactly unity requires an incredible fine-tuning. It is the task of the next sections to try to explain this.

7.2 'Old' and 'New' Inflation

The earliest time after the Big Bang we can meaningfully consider is Planck time t_P (Equation (6.65)), because earlier than that the theory of gravitation must be married to quantum field theory, a task which has not yet been mastered. Let us assume that the r_P-sized universe then was pervaded by a homogeneous scalar classical field φ, the *inflaton* field, and that all points in this universe were causally connected. The idea with inflation is to provide a mechanism which blows up the Universe so rapidly, and to such an enormous scale, that the causal connection between its different parts is lost, yet they are similar due to their common origin. This should solve the horizon problem and dilute the monopole density to acceptable values, as well as flatten the local fluctuations to near homogeneity.

Guth's Scenario. Let us try to make this idea more quantitative. Suppose that the mass m_φ of the inflaton carrying the field φ was much lighter than the Planck mass m_P,

$$0 < m_\varphi \ll m_P, \tag{7.22}$$

so that the inflaton can be considered to be massless. In fact, the particle symmetry at Planck time is characterized by all fields except the inflaton field being exactly massless. Only when this symmetry is spontaneously broken in the transition to a lower temperature phase do some particles become massive. We met this situation in Sections 6.5 and 6.6, where scalar Higgs fields played an important role.

Let us introduce the potential $V(\varphi, T)$ of the scalar field at temperature T. Its φ dependence is arbitrary, but we could take it to be a power function of φ just as we chose to do in Equations (4.73), (6.49) and (6.58). Suppose that the potential at time t_{P} has a minimum at a particular value φ_{P}. The Universe would then settle in that minimum, given enough time, and the value φ_{P} would gradually pervade all of space-time. It would be difficult to observe such a constant field because it would have the same value to all observers, regardless of their frame of motion. Thus the value of the potential $V(\varphi_{\mathrm{P}}, T_{\mathrm{P}})$ may be considered as a property of the vacuum.

Suppose that the minimum of the potential is at $\varphi_{\mathrm{P}} = 0$ in some region of space-time, and it is nonvanishing,

$$|V(0, T_{\mathrm{P}})| > 0. \tag{7.23}$$

An observer moving along a trajectory in space-time would notice that the field fluctuates around its *vacuum expectation value*

$$\langle \varphi_{\mathrm{P}} \rangle = 0,$$

and the potential energy consequently fluctuates around the mean vacuum-energy value

$$\langle V(0, T_{\mathrm{P}}) \rangle > 0.$$

This vacuum energy contributes a repulsive energy density to the total energy density in Friedmann's equation (4.17), acting just as dark energy or as a cosmological constant if we make the identification

$$\tfrac{1}{3} 8\pi G \langle V_0 \rangle \equiv \tfrac{1}{3}\lambda, \tag{7.24}$$

where $V_0 = V(0, 0)$ is a temperature-independent constant.

Inflation occurs when the Universe is dominated by the inflaton field φ and obeys the slow-roll conditions (4.74). We shall restrict our considerations to theories with a single inflaton field. Inflationary models assume that there is a moment when this domination starts and subsequently drives the Universe into a de Sitter-like exponential expansion in which $T \simeq 0$. *Alan Guth* in 1981 [2] named this an *inflationary universe*.

The timescale for inflation is

$$H = \sqrt{\frac{8\pi G}{3} \langle V_0 \rangle} \propto \frac{\sqrt{\hbar c}}{M_{\mathrm{P}}} \simeq (10^{-34}\ \mathrm{s})^{-1}. \tag{7.25}$$

Clearly the cosmic inflation cannot go on forever if we want to arrive at our present slowly expanding Friedmann–Lemaître universe. Thus there must be a

mechanism to halt the exponential expansion, a *graceful exit*. The freedom we have to arrange this is in the choice of the potential function $V(\varphi, T)$ at different temperatures T.

GUT Potentials. Suppose that there is a symmetry breaking phase transition from a hot $G_{\rm GUT}$-symmetric phase dominated by the scalar field φ to a cooler $G_{\rm s}$-symmetric phase. As the Universe cools through the critical temperature $T_{\rm GUT}$, bubbles of the cool phase start to appear and begin to grow. If the rate of bubble nucleation is initially small the Universe supercools in the hot phase, very much like a supercooled liquid which has a state of lowest potential energy as a solid.

We assumed above that the potential $V(\varphi, T_{\rm hot})$ in the hot $G_{\rm GUT}$-symmetric phase was symmetric around the point $\varphi = 0$, as in the top curve of Figure 6.7. Suppose now that $V(\varphi, T)$ develops a new asymmetric minimum as the temperature decreases. At $T_{\rm GUT}$, this second minimum may be located at $\varphi_{\rm GUT}$, and the potential may have become equally deep in both minima, as in the middle curve in Figure 6.7. Now the Universe could in principle *tunnel* out to the second minimum, but the potential barrier between the minima makes this process slow. Tunnelling through potential barriers is classically forbidden, but possible in quantum physics because quantum laws are probabilistic. If the potential barrier is high or wide, tunnelling is less probable. This is the reason why the initial bubble nucleation can be considered to be slow.

The lowest curve in Figure 6.7 illustrates the final situation when the true minimum has stabilized at φ_0 and the potential energy of this true vacuum is lower than in the original *false vacuum*:

$$V(\varphi_0, T_{\rm cool}) < V(0, T_{\rm hot}).$$

When the phase transition from the supercooled hot phase to the cool phase finally occurs at $T_{\rm cool}$ the latent heat stored as vacuum energy is liberated in the form of radiation and kinetic energy of ultrarelativistic massive scalar particles with positive pressure. At the same time other GUT fields present massify in the process of spontaneous symmetry breaking, suddenly filling the Universe with particles of temperature $T_{\rm R}$. The liberated energy is of the order of

$$\langle V_0 \rangle \simeq (kT_{\rm R})^4. \tag{7.26}$$

This heats the Universe enormously, from an ambient temperature

$$T_{\rm cool} \ll T_{\rm GUT}$$

to $T_{\rm R}$, which is at the $T_{\rm GUT}$ scale. The remaining energy in the φ field is dumped in the entropy which is proportional to T^3. Thus the entropy per particle is suddenly increased by the very large factor

$$Z^3 = \left(\frac{T_{\rm R}}{T_{\rm cool}}\right)^3, \tag{7.27}$$

where the ratio $T_{\rm R}/T_{\rm cool}$ is of the order of magnitude of 10^{29}. This is a nonadiabatic process, grossly violating the condition (5.13).

At the end of inflation the Universe is a hot bubble of particles and radiation in thermal equilibrium. The energy density term in Friedmann's equations has become dominant, and the Universe henceforth follows a Friedmann–Lemaître type evolution.

The flatness problem is now solved if the part of the Universe which became our Universe was originally homogeneous and has expanded by the de Sitter scale factor (4.60)

$$a = \mathrm{e}^{H\tau} \simeq 10^{29}, \tag{7.28}$$

or $H\tau \simeq 65$. Superimposed on the homogeneity of the pre-inflationary universe there were small perturbations in the field φ or in the vacuum energy. At the end of inflation these give rise to density perturbations which are the seeds of later mass structures and which can easily explain 10^{90} particles in the Universe.

It follows from Equations (7.25) and (7.28) that the duration of the inflation was

$$\tau \simeq 65 \times 10^{-34} \text{ s}. \tag{7.29}$$

Then also the horizon problem is solved, since the initial particle horizon has been blown up by a factor of 10^{29} to a size vastly larger than our present Universe. (Note that the realistic particle horizon is not infinite as one would obtain from Equation (7.9), because the lower limit of the integral is small but nonzero.) Consequently, all the large-scale structures seen today have their common origin in a microscopic part of the Universe long before the last scattering of radiation. The development of Guth's scenario through the pre-inflationary, inflationary and post-inflationary eras is similar to Linde's scenario shown in Figure 7.2, except that the vertical scale here grows 'only' to 10^{29}.

When our bubble of space-time nucleated, it was separated from the surrounding supercooled hot phase by domain walls. When the phase transition finally occurred the enormous amounts of latent heat was released to these walls. The mechanism whereby this heat was transferred to particles in the bubbles was by the collision of domain walls and the coalescence of bubbles. In some models knots or topological defects then remained in the form of monopoles of the order of one per bubble. Thus the inflationary model also solves the monopole problem by blowing up the size of the region required by one monopole. There remains no inconsistency then with the present observed lack of monopoles.

Although Guth's classical model of cosmic inflation may seem to solve all the problems of the hot Big Bang scenario in principle, it still fails because of difficulties with the nucleation rate. If the probability of bubble formation is large, the bubbles collide and make the Universe inhomogeneous to a much higher degree than observed. If the probability of bubble formation is small, then they never collide and there is no reheating in the Universe, so each bubble remains empty. Thus there is no graceful exit from the inflationary scenario.

The first amendments to Guth's model tried to modify the φ^4-potential of Equations (6.49) and (6.58) in such a way that the roll from the false minimum to the true minimum would start very slowly, and that the barrier would be very small or absent. This model has been called *new inflation* by its inventors, A. D. Linde [3, 4, 5] and A. Albrecht and P. J. Steinhardt [6]. However, this required unlikely

fine-tuning, and the amplitude of primordial quantum fluctuations needs to be smaller than $10^{-6}M_{\rm P}$ in order not to produce too large late-time density fluctuations.

7.3 Chaotic Inflation

Initial Conditions. Guth's model made the rather specific assumption that the Universe started out with the vacuum energy in the false minimum $\varphi = 0$ at time $t_{\rm P}$. However, Linde pointed out that this value as well, as any other fixed starting value, is as improbable as complete homogeneity and isotropy because of the quantum fluctuations at Planck time (see, for example, references [7, 8]). Instead, the scalar field may have had some random starting value φ_a, which could be assumed to be fairly uniform across the horizon of size $M_{\rm P}^{-1}$, changing only by an amount

$$\Delta\varphi_a \simeq M_{\rm p} \ll \varphi_a. \tag{7.30}$$

Regions of higher potential would expand faster, and come to dominate. With time the value of the field would change slowly until it finally reached φ_0 at the true minimum $V(\varphi_0)$ of the potential.

But causally connected spaces are only of size $M_{\rm P}^{-1}$, so even the metric of spacetime may be fluctuating from open to closed in adjacent spaces of this size. Thus the Universe can be thought of as a chaotic foam of causally disconnected bubbles in which the initial conditions are different, and which would subsequently evolve into different kinds of universes. Only one bubble would become our Universe, and we could never get any information about the other ones. Linde called this essentially anthropic idea *chaotic inflation*.

According to Heisenberg's uncertainty relation, at a timescale $\Delta t = \hbar/M_{\rm P}c^2$ the energy is uncertain by an amount

$$\Delta E > \frac{\hbar}{\Delta t} = M_{\rm P}c^2. \tag{7.31}$$

Let us for convenience work in units common to particle physics where $\hbar = c = 1$. Then the energy density is uncertain by the amount

$$\Delta\rho = \frac{\Delta E}{(\Delta r)^3} = \frac{\Delta E}{(\Delta t)^3} = M_{\rm P}^4. \tag{7.32}$$

Thus there is no reason to assume that the potential $V(\varphi_a)$ would be much smaller than $M_{\rm P}^4$. We may choose a general parametrization for the potential,

$$V(\varphi) \approx \frac{\kappa\varphi^n}{nM_{\rm P}^{n-4}} \approx M_{\rm P}^4, \tag{7.33}$$

where $n > 0$ and $0 < \kappa \ll 1$. This assumption ensures that $V(\varphi_a)$ does not rise too steeply with φ. For $n = 4$ it then follows that

$$\varphi_a \simeq \left(\frac{4}{\kappa}\right)^{1/4} M_{\rm P} \gg M_{\rm P} \tag{7.34}$$

when the free parameter κ is chosen to be very small.

A large number of different models of inflation have been studied in the literature. Essentially they differ in their choice of potential function. We shall only study the simplest example of Equation (7.33) with $n = 2$.

Scalar-Field Dynamics. In the simplest field theory coupling a scalar field φ to gravitation, the total inflaton energy is of the form

$$\tfrac{1}{2}\dot{\varphi}^2 + \tfrac{1}{2}(\nabla\varphi)^2 + V(\varphi), \tag{7.35}$$

and the dynamics can be described by two equations: Friedmann's equation

$$H^2 + \frac{k}{a^2} = \frac{8\pi}{3M_{\mathrm{P}}^2}(\tfrac{1}{2}\dot{\varphi}^2 + \tfrac{1}{2}(\nabla\varphi)^2 + V(\varphi)); \tag{7.36}$$

and the Klein–Gordon equation (4.67) obeyed by scalar fields,

$$\ddot{\varphi} + 3H\dot{\varphi} + V'(\varphi) = 0. \tag{7.37}$$

If the inflaton field starts out as φ_a, being large and sufficiently homogeneous as we assumed in Equation (7.30), we have

$$(\nabla\varphi_a)^2 \ll V(\varphi_a). \tag{7.38}$$

The speed of the expansion, $H = \dot{a}/a$, is then dominated by the potential $V(\varphi_a)$ in Equation (7.36) and therefore large. (Note that we have also neglected all other types of energy, like ρ_{m} and ρ_{r}.)

Since the potential (7.33) has a minimum at $\varphi = 0$, one may expect that φ should oscillate near this minimum. However, in a rapidly expanding universe, the inflaton field approaches the minimum very slowly, like a ball in a viscous medium, the viscosity $V'(\varphi)$ being proportional to the speed of expansion. In this situation we have

$$\ddot{\varphi}_a \ll 3H\dot{\varphi}_a, \qquad \dot{\varphi}_a^2 \ll V(\varphi_a), \qquad \frac{k}{a^2} \ll H^2. \tag{7.39}$$

The first inequality states that φ changes so slowly that its acceleration can be neglected. The second inequality sets the condition for expansion: the kinetic energy is much less than the potential energy, so the pressure p_φ of the scalar field is negative and the expansion accelerates. In the expansion the scale factor a grows so large that the third inequality follows. The equations (7.36) and (7.37) then simplify to

$$H^2 = \frac{8\pi}{3M_{\mathrm{P}}^2}V(\varphi) \tag{7.40}$$

and

$$3H\dot{\varphi} = -V'(\varphi). \tag{7.41}$$

Equation (7.41) then describes an exponentially expanding de Sitter universe. Initially all space-time regions of size $H^{-1} = M_{\mathrm{P}}^{-1}$ would contain inhomogeneities inside their respective event horizons. At every instant during the inflationary

de Sitter stage an observer would see himself surrounded by a black hole with event horizon H^{-1} (but remember that 'black hole' really refers to a static metric). There is an analogy between the Hawking radiation of black holes and the temperature in an expanding de Sitter space. Black holes radiate at the Hawking temperature $T_{\rm H}$ (Equation (3.31)), while an observer in de Sitter space will feel as if he is in a thermal bath of temperature $T_{\rm dS} = H/2\pi$.

Within a time of the order of H^{-1} all inhomogeneities would have traversed the Hubble radius. Thus they would not affect the physics inside the de Sitter universe which would be getting increasingly homogeneous and flat. On the other hand, the Hubble radius is also receding exponentially, so if we want to achieve homogeneity it must not run away faster than the inhomogeneities.

Combining Equations (7.33), (7.40), and (7.41), we obtain an equation for the time dependence of the scalar field,

$$\tfrac{1}{2}\dot{\varphi}^2 = \frac{n^2 M_{\rm P}^2}{48\pi\varphi^2} V(\varphi). \tag{7.42}$$

Let us study the solution of this equation in the case when the potential is given by Equation (6.46) and Figure 6.4:

$$V(\varphi) = \tfrac{1}{2} m_\varphi^2 \varphi^2. \tag{7.43}$$

The time dependence of the field is then

$$\varphi(t) = \varphi_a - \frac{m_\varphi M_{\rm P}}{2\sqrt{3\pi}} t \equiv \varphi_a \left(\frac{1-t}{\tau}\right), \tag{7.44}$$

where $\tau(\phi_a)$ is the characteristic timescale of the expansion. At early times when $t \ll \tau$ the scalar field remains almost constant, changing only slowly from φ_a to its ultimate value φ_0. The scale factor then grows quasi-exponentially as

$$R(t) = R(t_a) \exp(Ht - \tfrac{1}{6} m_\varphi^2 t^2), \tag{7.45}$$

with H given by

$$H = 2\sqrt{\frac{\pi}{3}} \frac{m_\varphi}{M_{\rm P}} \varphi_a. \tag{7.46}$$

As the field approaches $\varphi_0 = 0$ the slow-roll of the Universe ends in the true vacuum at $V(\varphi_0)$, and the inflation ends in graceful exit.

Size of the Universe. At time τ, the Universe has expanded from a linear size $R(t_a)$ to

$$R(\tau) \simeq R(t_a) \exp(H\tau) = R(t_a) \exp\left(\frac{4\pi\varphi_a^2}{M_{\rm P}^2}\right). \tag{7.47}$$

For instance, a universe of linear size equal to the Planck length $R(t_a) \simeq 10^{-35}$ m has grown to

$$R(\tau) \simeq R(t_a) \exp\left(\frac{4\pi M_{\rm P}^2}{m_\varphi^2}\right). \tag{7.48}$$

For a numerical estimate we need a value for the mass m_φ of the inflaton. This is not known, but we can make use of the condition that the chaotic model must be able to form galaxies of the observed sizes. Then (as we shall see in the next chapters) the scalar mass must be of the order of magnitude

$$m_\varphi \simeq 10^{-6} M_{\mathrm{P}}. \tag{7.49}$$

Inserting this estimate into Equation (7.48) we obtain the completely unfathomable scale

$$R(\tau) \simeq 10^{-35\exp(4\pi\times 10^{12})}\ \mathrm{m} \simeq 10^{5.5\times 10^{12}}\ \mathrm{m}. \tag{7.50}$$

It is clear that all the problems of the standard Big Bang model discussed in Section 7.1 then disappear. The homogeneity, flatness and isotropy of the Universe turn out to be consequences of the inflaton field having been large enough in a region of size M_{P}^{-1} at time t_{P}. The inflation started in different causally connected regions of space-time 'simultaneously' to within 10^{-43} s, and it ended at about 10^{-35} s. Our part of that region was extremely small. Since the curvature term in Friedmann's equations decreased exponentially, the end result is exactly as if k had been zero to start with. A picture of this scenario is shown in Figure 7.2.

Quantum Fluctuations. If inflation was driven by a pure de Sitter expansion, the enormous scale (7.50) guarantees that it would henceforth be absolutely flat. But we noted that at Planck time the field φ was indefinite by M_{p}, at least, so that there were deviations from a pure de Sitter universe. Even if this Universe was empty, quantum field theory tells us that empty space is filled with zero-point quantum fluctuations of all kinds of physical fields, here fluctuations from the classical de Sitter inflaton field.

The vacuum fluctuation spectrum of the slowly rolling scalar field during the inflationary expansion turns out to be quite unlike the usual spectrum of thermal fluctuations. This can be seen if one transforms the de Sitter metric (4.61) into the metric of a Euclidean 4-sphere (2.27). Bose fields (like the inflaton) obeying a massless Klein–Gordon equation turn out to oscillate harmonically on this sphere with period $2\pi/H$, which is equivalent to considering quantum statistics at a temperature $T_{\mathrm{dS}} = H/2\pi$. However, the temperature in de Sitter space is highly unusual in that the fluctuations on the 4-sphere are periodic in all four dimensions [7, 8].

The fate of a bubble of space-time clearly depends on the value of φ_a. Only when it is large enough will inflationary expansion commence. If φ_a is very much larger than M_{P}, Equation (7.46) shows that the rate of expansion is faster than the timescale τ,

$$H \gg 2\sqrt{\frac{\pi}{3}} m_\varphi \simeq \frac{2}{\tau}. \tag{7.51}$$

Although the wavelengths of all quantum fields then grow exponentially, the change $\Delta\varphi$ in the value of the inflaton field itself may be small. In fact, when the physical wavelengths have reached the size of the Hubble radius H^{-1}, all changes in φ are impeded by the friction $3H\dot{\varphi}$ in Equation (7.41), and fluctuations of size

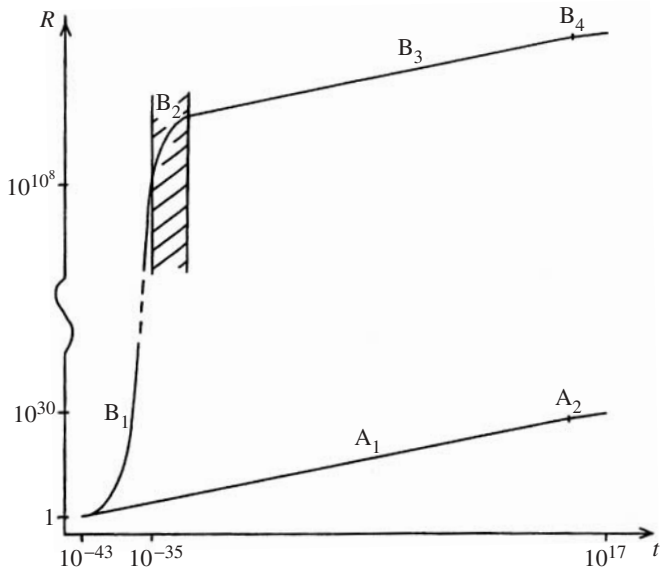


Figure 7.2 Evolution of the scale R of the Universe since Planck time in (A) Friedmann models and (B) inflationary expansion. During the epoch B_1 the Universe expands exponentially, and during B_2 the inflation ends by reheating the Universe. After the graceful exit from inflation the Universe is radiation dominated along B_3, just as in A_1, following a Friedmann expansion. The sections B_4 and A_2 are matter-dominated epochs.

$\delta\varphi$ freeze to an average nonvanishing amplitude of

$$|\delta\varphi(x)| \simeq \frac{H}{2\pi}. \tag{7.52}$$

Consequently, the vacuum no longer appears empty and devoid of properties.

Fluctuations of a length scale exceeding H^{-1} are outside the present causal horizon so they no longer communicate, crests and troughs in each oscillation mode remain frozen. But at the end of inflation, the expansion during radiation and matter domination returns these frozen fluctuations inside the horizon. With time they become the seeds of perturbations we now should observe in the CMB and in the density distribution of matter.

The quantum fluctuations remaining in the inflaton field will cause the energy to be dumped into entropy at slightly fluctuating times. Thus, the Universe will also contain entropy fluctuations as seeds of later density perturbations.

Linde's Bubble Universe. Since our part of the pre-inflationary universe was so small, it may be considered as just one bubble in a foam of bubbles having different fates. In Linde's chaotic model each original bubble has grown in one e-folding time $\tau = H^{-1}$ to a size comprising e^3 mini-universes, each of diameter H^{-1}. In half of these mini-universes, on average, the value of φ may be large enough for inflation to continue, and in one-half it may be too small. In the next e-folding time the same pattern is repeated. Linde has shown that in those parts

of space-time where φ grows continuously the volume of space grows by a factor

$$e^{(3-\ln 2)Ht}, \tag{7.53}$$

whereas in the parts of space-time where φ does not decrease the volume grows by the factor

$$\tfrac{1}{2}e^{3Ht}. \tag{7.54}$$

Since the Hubble parameter is proportional to φ, most of the physical volume must come from bubbles in which φ is maximal:

$$\varphi \simeq M_P^2/m_\varphi. \tag{7.55}$$

But there must also be an exponential number of bubbles in which φ is smaller. Those bubbles are the possible progenitors of universes of our kind. In them, φ attains finally the value corresponding to the true minimum $V(\varphi_0)$, and a Friedmann–Lemaître-type evolution takes over. Elsewhere the inflatoric growth continues forever. Thus we happen to live in a Universe which is a minuscule part of a steady-state eternally inflating meta-Universe which has no end, and therefore it also has no beginning. There is simply no need to turn inflation on in the first place, and the singularity at time zero has dropped out from the theory.

During inflation, each bubble is generating new space-time to expand into, as required by general relativity, rather than expanding into pre-existing space-time. In these de Sitter space-times the bubble wall appears to an observer as a surrounding black hole. Two such expanding bubbles are causally disconnected, so they can neither collide nor coalesce. Thus the mechanism of vacuum-energy release and transfer of heat to the particles created in the phase transition is not by bubble collisions as in the classical model. Instead, the rapid oscillations of the inflaton field φ decay away by particle production as the Universe settles in the true minimum. The potential energy then thermalizes and the Universe reheats to some temperature of the order of T_{GUT}.

In this reheating, any baryon–anti-baryon asymmetry produced during the GUT phase transition mechanism is washed out, that is why some other phase transition must be sought to explain the baryon–anti-baryon asymmetry. Thus the existence of baryons is an indication that particle physics indeed has to go beyond the 'minimal standard model'.

Predictions. One consequence of the repulsive scalar field is that any two particles appear to repel each other. This is the Hubble expansion, which is a consequence of inflation. In noninflationary theories the Hubble expansion is merely taken for granted.

Inflationary models predict that the density of the Universe should today be critical,

$$\Omega_0 = 1. \tag{7.56}$$

Consequently, we should not only observe that there is too little luminous matter to explain the dynamical behaviour of the Universe, we also have a precise

theoretical specification for how much matter there should be. This links dark matter to inflation.

We have already noted that the scalar inflaton field produced a spectrum of frozen density and radiation perturbations beyond the horizon, which moved into sight when the expansion of the Universe decelerated. In the post-inflationary epoch when the Friedmann expansion takes over we can distinguish between two types of perturbations, or perturbations and or perturbations. In the first case, the perturbations in the local number density, $\delta_{\rm m} \equiv \delta\rho_{\rm m}/\rho_{\rm m}$, of each species of matter—baryons, leptons, neutrinos, dark matter—is the same. In particular, these perturbations are coupled to those of radiation, $\delta_{\rm r} \equiv \delta\rho_{\rm r}/\rho_{\rm r}$, so that $4\delta_{\rm m} = 3\delta_{\rm r}$ (from Equation (5.45)). By the principle of covariance, perturbations in the energy-momentum tensor imply simultaneous perturbations in energy density and pressure, and by the equivalence principle, variations in the energy-momentum tensor are equivalent to variations in the curvature. Curvature perturbations can have been produced early as irregularities in the metric, and they can then have been blown up by inflation far beyond the Hubble radius. Thus adiabatic perturbations are a natural consequence of cosmic inflation. In contrast, inflation does not predict any isocurvature perturbations.

Let us write the power spectrum of density perturbations in the form

$$P(k) \propto k^{n_{\rm s}}, \tag{7.57}$$

where $n_{\rm s}$ is the *scalar spectral index*. Inflationary models predict that the primordial fluctuations have an equal amplitude on all scales, an almost scale-invariant power spectrum as the matter fluctuations cross the Hubble radius, and are Gaussian. This is the Harrison–Zel'dovich spectrum for which $n_{\rm s} = 1$ ($n_{\rm s} = 0$ would correspond to white noise).

A further prediction of inflationary models is that tensor fluctuations in the space-time metric, satisfying a massless Klein–Gordon equation, have a nearly scale-invariant spectrum of the form (7.57) with *tensor spectral index* $n_{\rm t} \approx 1$, just like scalar density perturbations, but independently of them.

The above predictions are generic for a majority if inflation models which differ in details. Inflation as such cannot be either proved or disproved, but specific theories can be and have been ruled out. In Chapters 8 and 9 we shall test the validity of the above predictions.

7.4 The Inflaton as Quintessence

Now we have met two cases of scalar fields causing expansion: the inflaton field acting before $t_{\rm GUT}$ and the quintessence field describing present-day dark energy. It would seem economical if one and the same scalar field could do both jobs. Then the inflaton field and quintessence would have to be matched at some time later than $t_{\rm GUT}$. This seems quite feasible since, on the one hand, the initially dominating inflaton potential $V(\varphi)$ must give way to the background energy density $\rho_{\rm r} + \rho_{\rm m}$ as the Universe cools, and on the other hand, the dark energy density must have been much smaller than the background energy density until recently.

Recall that quintessence models are constructed to be quite insensitive to the initial conditions.

On the other hand, nothing forces the identification of the inflaton and quintessence fields. The inflationary paradigm in no way needs nor predicts quintessence.

In the previously described models of inflation, the inflaton field φ settled to oscillate around the minimum $V(\varphi = 0)$ at the end of inflation. Now we want the inflaton energy density to continue a monotonic roll-down toward zero, turning ultimately into a minute but nonvanishing quintessence tail. The global minimum of the potential is only reached in a distant future, $V(\varphi \to \infty) \to 0$. In this process the inflaton does not decay into a thermal bath of ordinary matter and radiation because it does not interact with particles at all, it is said to be sterile. A sterile inflaton field avoids violation of the equivalence principle, otherwise the interaction of the ultralight quintessence field would correspond to a new long-range force. Entropy in the matter fields comes from gravitational generation at the end of inflation rather than from decay of the inflaton field.

The task is then to find a potential $V(\varphi)$ such that it has two phases of accelerated expansion: from $t_{\rm P}$ to $t_{\rm end}$ at the end of inflation, and from a time $t_{\rm F} \approx t_{\rm GUT}$ when the instanton field freezes to a constant value until now, t_0. Moreover, the inflaton energy density must decrease faster than the background energy density, equalling it at some time t_* when the field is φ_*, and thereafter remaining subdominant to the energy density of the particles produced at $t_{\rm end}$. Finally it must catch up with a tracking potential at some time during matter domination, $t > t_{\rm eq}$.

The mathematical form of candidate potentials is of course very complicated, and it would not be very useful to give many examples here. However, it is instructive to follow through the physics requirements on φ and $V(\varphi)$ from inflation to present.

Kination. Inflation is caused by an essentially constant potential $V(\varphi)$ according to Equation (7.40). The condition $V(\varphi \to \infty) \to 0$ requires an end to inflation at some finite time $t_{\rm end}$ when the field is $\varphi_{\rm end}$ and the potential is $V_{\rm end} \equiv V(\varphi_{\rm end})$. The change in the potential at $t_{\rm end}$ from a constant to a decreasing roll then implies, by Equation (7.41), that $\dot\varphi_{\rm end} \neq 0$, and furthermore, by Equation (7.37), that also $\ddot\varphi_{\rm end} \neq 0$. Then the slow-roll conditions (4.74) for ϵ and η are also violated.

During inflation the kinetic energy density of the inflaton is

$$\rho_{\rm kin} = \epsilon V = \frac{m_{\rm Planck}^2}{16\pi}\left(\frac{V'^2(\varphi)}{V(\varphi)}\right). \tag{7.58}$$

Thus when $V'(\varphi)$ starts to grow, so does $\rho_{\rm kin}$. and the total energy density of the Universe becomes dominated by the inflaton kinetic energy density. This epoch has been called *kination* or *deflation*. Equation (4.70) then dictates that the equation of state is

$$w_\varphi = \frac{\dot\varphi^2 + 2V(\varphi)}{\dot\varphi^2 - 2V(\varphi)} \approx 1, \tag{7.59}$$

so that the kinetic energy density decreases as

$$\rho(a) \propto a^{-3(1+w)} = a^{-6} \tag{7.60}$$

from Equation (4.29). This is much faster than the a^{-4} decrease of the radiation energy density $\rho_{\rm r}$, and the a^{-3} decrease of the initially much smaller matter energy density $\rho_{\rm m}$. Consequently, kination ends at the moment when $\rho_{\rm r}$ overtakes $\rho_{\rm kin}$ at time t_*. When constructing phenomenological models for this scenario, one constraint is of course that $\rho_{\rm r}(t_{\rm end}) \ll V_{\rm end}$, or equivalently, $t_{\rm end} < t_*$. This behaviour is well illustrated in Figure 7.3, taken from the work of Dimopoulos and Valle [9].

Since matter and radiation are gravitationally generated at $t_{\rm end}$, the reheating temperature of radiation is given by

$$T_{\rm reh} = \alpha T_{\rm H}, \tag{7.61}$$

where $T_{\rm H}$ is the Hawking temperature (Equation (3.31)), and α is some reheating efficiency factor less than unity. In Figure 7.3 the radiation energy density $\rho_\gamma \equiv \rho_{\rm r}$ starts at $T_{\rm reh}^4 \ll V_{\rm end}$, and then catches up with $V(\varphi)$ at φ_*. Now the Universe becomes radiation dominated and the hot Big Bang commences. Note that the term 'hot Big Bang' has a different meaning here: it does not refer to a time zero with infinite temperature, but to a moment of explosive entropy generation. This mimics the Big Bang so that all of its associated successful predictions ensue.

Quintessence. The inflaton field plays no role any more during the Big Bang. The kinetic energy density reduces rapidly to negligible values by its a^{-6} dependence and the field freezes ultimately to a nonzero value $\varphi_{\rm F}$. The residual inflaton potential $V(\varphi)$ again starts to dominate over the kinetic energy density, however, staying far below the radiation energy density and, after $t_{\rm eq}$, also below the matter energy density.

As we approach t_0, the task of phenomenology is to devise a quintessence potential having a suitable tracker. The nature of the tracker potential is decided by the form of the quintessence potential. To arrive at the present-day dark energy which causes the evolution to accelerate, the field φ must be unfrozen again, so $\phi_{\rm F}$ should not be very different from ϕ_0. Many studies have concluded that only exponential trackers are admissible, and that quintessence potentials can be constructed by functions which behave as exponentials in φ early on, but which behave more like inverse power potentials in the quintessential tail. A simple example of such a potential is

$$V(\varphi \gg \varphi_{\rm end}) \approx V_{\rm end} \frac{\exp(-\lambda\varphi/m_{\rm P})}{(\varphi/m)^k}, \tag{7.62}$$

where $k \geqslant 1$ is an integer, $\lambda > 0$ is a parameter characterizing the exponential scale, and $m < m_{\rm P}$ is a mass scale characteristic of the adopted inverse power-law scale. For more details, see Dimopoulos and Valle [9].

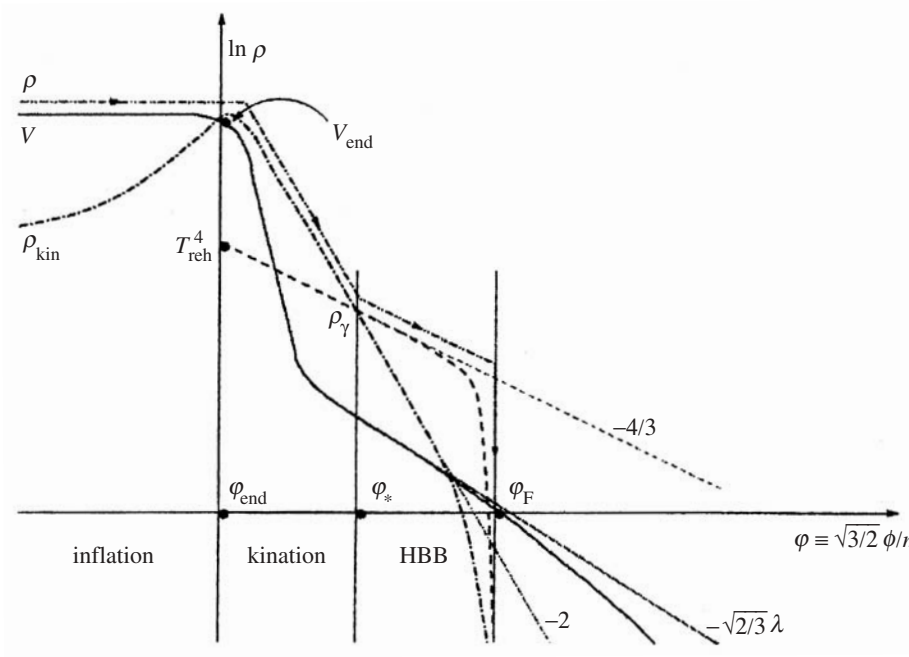


Figure 7.3 Schematic view of the scalar potential from inflation to quintessence. The potential V (solid line) features two flat regions, the inflationary plateau and the quintessential tail. The inflation is terminated at $\varphi_{\rm end}$ by a drastic reduction of V, leading to a rapid roll-down of the scalar field from the inflationary plateau towards the quintessential tail. At the end of inflation the kinetic energy density of the scalar field, $\rho_{\rm kin}$ (dash-dotted line), dominates for a brief period the energy density of the Universe. During this time the radiation energy density ρ_y (dashed line) reduces less rapidly and catches up with $\rho_{\rm kin}$ at time t_* when the field is φ_*, and the explosive generation of entropy commences. After that the kinetic energy of the scalar field reduces rapidly to zero and the field freezes asymptotically to a value $\varphi_{\rm F}$, while the overall energy density of the Universe (dash-dot-dotted line) continues to decrease due to the Hubble expansion. Assuming a quasi-exponential tail given by Equation (7.62), the potential beyond $\varphi_{\rm F}$ is seen departing logarithmically from a pure exponential case (dotted line). Reprinted from K. Dimopoulos and J. W. F. Valle, *Modeling quintessential inflation* [9], copyright 2002, with permission from Elsevier.

7.5 Cyclic Models

As we have seen, 'consensus' inflation by a single inflaton field solves the problems described in Section 7.1. But in the minds of some people it does so at a very high price. It does not explain the beginning of space and time, it does not predict the future of the Universe, or it sweeps these fundamental questions under the carpet of the Anthropic Principle. It invokes several unproven ingredients, such as a scalar field and a scalar potential, suitably chosen for the field to slow-roll down the potential while its kinetic energy is negligible, and such that it comes to a graceful exit where ordinary matter and radiation are created by oscillations in the potential well, or by entropy generation during a second slow-roll phase of an

equally arbitrary dark energy field. Clearly, any viable alternative to single-field inflation must also be able to solve the problems in Section 7.1, and it should not contain more arbitrary elements than does single-field inflation—multiple scalar fields have more freedom but also more arbitrariness.

The Entropy Problem. In the radiation-dominated Universe, the source of energy of photons and other particles is a phase transition or a particle decay or an annihilation reaction, many of these sources producing monoenergetic particles. Thus the energy spectrum produced at the source is very nonuniform and nonrandom, containing sharp spectral lines 'ordered' by the type of reaction. Such a spectrum corresponds to low entropy. Subsequent scattering collisions will redistribute the energy more randomly and ultimately degrade it to high-entropy heat. Thermal equilibrium is the state of maximum uniformity and highest entropy. The very fact that thermal equilibrium is achieved at some time tells us that the Universe must have originated in a state of low entropy.

In the transition from radiation domination to matter domination no entropy is lost. We have seen the crucial effect of photon reheating due to entropy conservation in the decoupling of the electrons. As the Universe expands and the wavelengths of the CMB photons grow, the available energy is continuously converted into lower-grade heat, thus increasing entropy. This thermodynamic argument defines a *preferred direction of time*.

When the cooling matter starts to cluster and contract under gravity, a new phase starts. We have seen that the Friedmann–Lemaître equations dictate instability, the lumpiness of matter increases with the Universe developing from an ordered, homogeneous state towards chaos. It may seem that contracting gas clouds represent higher uniformity than matter clustered into stars and galaxies. If the only form of energy were thermal, this would indeed be so. It would then be highly improbable to find a gas cloud squeezed into a small volume if nothing hinders it from filling a large volume. However, the attractive nature of gravity seemingly reverses this logic: the cloud gains enormously in entropy by contracting. Thus the preferred direction of time as defined by the direction of increasing entropy is unchanged during the matter-dominated era.

The same trend continues in the evolution of stars. Young suns burn their fuel through a chain of fusion reactions in which energetic photons are liberated and heavier nuclei are produced. As the photons diffuse in the stellar matter, they ultimately convert their energy into a large number of low-energy photons and heat, thereby increasing entropy. Old suns may be extended low-density, high-entropy red giants or white dwarfs, without enveloping matter which loses mass by various processes. In the process of supernova explosion entropy grows enormously.

Consider the contracting phase of an oscillating universe. After the time $t_{\max}$ given by Equation (4.52) the expansion turns into contraction, and the density of matter grows. If the age of the Universe is short enough that it contains black holes which have not evaporated, they will start to coalesce at an increasing rate. Thus entropy continues to increase, so that the preferred direction of time is

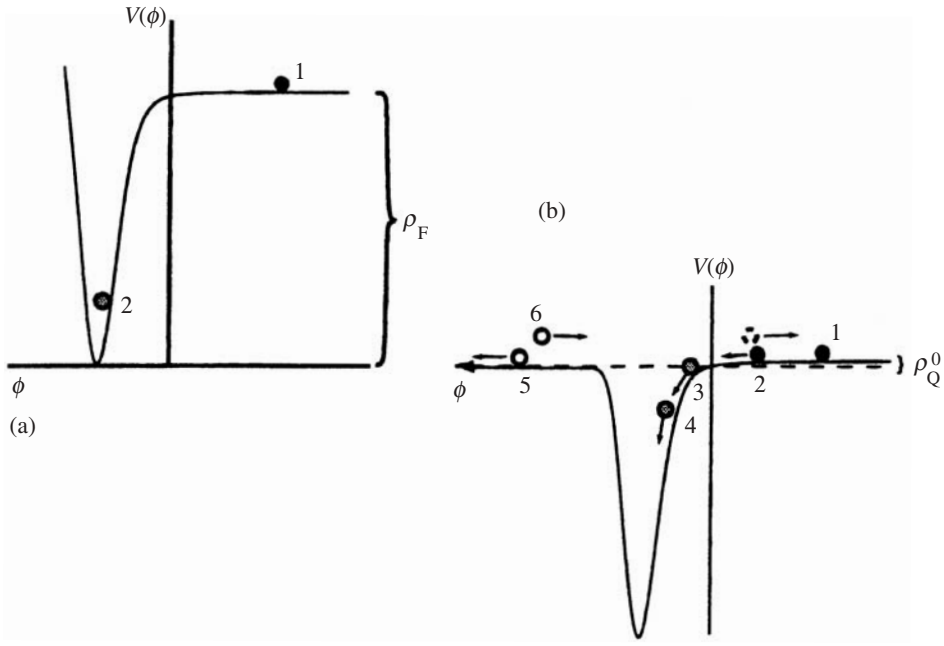


Figure 7.4 Schematic view of the potential $V(\varphi)$ as a function of the field φ for (a) inflationary cosmology and (b) cyclic models. The numbered sequence of stages is described in the text. From ref. [10] courtesy of P. J. Steinhardt.

unchanged. Shortly before the Big Crunch, when the horizon has shrunk to linear size $L_{\rm p}$, all matter has probably been swallowed by one enormously massive black hole.

A Cyclic Universe. Early attempts to build models with cyclically reoccurring expansion and contraction were plagued by this problem, that the entropy density would rise from cycle to cycle. The length of cycles must then increase steadily. But, in retrospect, there must then have been a first cycle a finite time ago, thus a beginning of time: precisely what the cyclic model was conceived to avoid.

A cyclic model which solves the entropy problem and which appears as successful as the 'consensus' inflationary model (leaving aside whether this comes at a higher or lower price) has been proposed by Steinhardt, Turok and collaborators [10]. The model is described qualitatively in Figure 7.4, which depicts a potential $V(\varphi)$, function of a scalar field φ. Unlike the inflaton field, φ does not cause an inflationary expansion of space-time.

Each cycle ends and begins with crunch turning into a bang at the field value $\varphi = -\infty$. The bang is a transition or bounce from a pre-existing contracting phase with a with a decreasing field, into an expanding phase with an increasing field. The contraction occurs in the extra dimension φ, rather than in our three dimensions. In the acceleration of the field at turn-around, matter and radiation are created

at large but finite temperature from the kinetic energy of the field (stage 6 in Figure 7.4).

The bang is then followed by an immediate entry into a period of radiation and matter domination where the field is rushing towards positive values (stage 7). This stage is quite similar to the corresponding post-inflationary epoch in the conventional inflationary scenario, and therefore the predictions are the same. But, unlike the conventional model, here a subdominant dark energy field is required. During radiation and matter domination the scalar dark energy field is effectively frozen in place by the Hubble redshift of its kinetic energy. The potential energy of this field starts to dominate only when radiation and matter energy densities have been sufficiently diluted by the expansion; then a slow cosmic acceleration commences (stage 1), and a slow roll down the weakly sloping potential (stage 2).

At this stage the potential energy of the scalar field dominates also over the kinetic energy of the scalar field so that dark energy drives the expansion, much like the post-inflationary quintessence in the previous section. Next the field crosses zero potential (stage 3) and the kinetic energy starts to dominate over the now negative potential, causing the expansion to stop and turn into contraction with equation of state $w \gg 1$.

The potential has a minimum, where it is strongly negative. As the scalar field rolls down into this minimum (stage 4), it picks up speed on the way and causes a Hubble blueshift of the scalar field kinetic energy. With this speed the field just continues beyond the minimum, climbs out of it on the other side and continues towards negative infinity in a finite time.

In the approach towards $\varphi = -\infty$, the scale factor $a(t)$ goes to zero (stage 5), but the coupling of the scalar field to radiation and matter conspires in such a way as to keep the temperature and the energy density finite. In order for space not to disappear completely, this scenario has to take place in a five-dimensional space-time, and $a(t)$ has to be interpreted as the effective scale factor on a four-dimensional *brane* of the full five-dimensional space-time.

The homogeneity and flatness of the Universe and the density perturbations are established during long periods of ultra-slow accelerated expansion, and the conditions are set up during the negative time prior to a bang. In contrast, inflationary theories have very little time to set up large-scale conditions, only about 10^{-35} s until the inflationary fluctuations have been amplified and frozen in, at a length scale of 10^{-25}cm. In the cyclic Universe, the fluctuations can be generated a fraction of a second before the bang when their length scale is thousands of kilometres.

Although the acceleration due to dark energy is very slow, causing the Universe to double in size every 15 billion years or so, compared with the enormous expansion in Equation (7.50) during 10^{-35} s, this is enough to empty the Universe of its matter and radiation. The dark energy dilutes the entropy density to negligible levels at the end of each cycle (at stage 1), preparing the way for a new cycle of identical duration. Although the total entropy of the Universe and the number of black holes increase from cycle to cycle, and increase per comoving volume as well, the physical entropy density and the number of black holes per proper vol-

ume is expanded away in each cycle. And since it is the physical entropy density that determines the expansion rate, the expansion and contraction history is the same from cycle to cycle.

All of this is very speculative, but so is consensus inflation and quintessence dark energy. Fortunately, the different models make different testable predictions, notably for gravitational radiation.

Problems

1. Derive Equations (7.41) and (7.42).
2. Derive $\varphi(t)$ for a potential $V(\varphi) = \frac{1}{4}\lambda\varphi^4$.
3. Suppose that the scalar field averaged over the Hubble radius H^{-1} fluctuates by an amount ψ. The field gradient in this fluctuation is $\nabla\psi = H\psi$ and the gradient energy density is $H^2\psi^2$. What is the energy of this fluctuation integrated over the Hubble volume? Use the timescale H^{-1} for the fluctuation to change across the volume and the uncertainty principle to derive the minimum value of the energy. This is the amount by which the fluctuation has stretched in one expansion time [11].
4. Material observed now at redshift $z = 1$ is at present distance H_0^{-1}. The recession velocity of an object at coordinate distance x is $\dot{R}x$. Show that the recession velocity at the end of inflation is
$$\dot{R}x = \frac{H_0 R_0 x z_{\rm r}}{\sqrt{z_{\rm eq}}}, \tag{7.63}$$
where $z_{\rm r}$ is the redshift at the end of inflation. Compute this velocity. The density contrast has grown by the factor $z_{\rm r}^2/z_{\rm eq}$. What value did it have at the end of inflation since it is now $\delta \approx 10^{-4}$ at the Hubble radius [11]?
5. Show that in an exponentially expanding universe ($q = -1$) the Hubble sphere is stationary. Show that it constitutes an event horizon in the sense that events beyond it will never be observable. Show that in this universe there is no particle horizon [12].
6. Show that the number of e-foldings of inflation in the $V(\varphi) = -\lambda\varphi^4$ model is of order
$$N \approx \frac{H^2}{\lambda\varphi_i^2}$$
from the time at which the field has the value φ_i to the end of inflation ($\varphi \ll \varphi_i$). Hence show that the density perturbations in this model are of order
$$\left(\frac{\delta\rho}{\rho}\right)_H \approx \sqrt{\lambda N^3}. \tag{7.64}$$
Deduce that $\lambda < 10^{-14}$ is required if the fluctuations are to be compatible with the CMB. This of course amounts to the fine-tuning that inflation is supposed to avoid [12].

Chapter Bibliography

[1] Kaiser, N. and Silk, J. 1986 *Nature* **324**, 529.
[2] Guth, A. H. 1981 *Phys. Rev.* D **23**, 347.
[3] Linde, A. D. 1982 *Phys. Lett.* B **108**, 389.
[4] Linde, A. D. 1982 *Phys. Lett.* B **114**, 431.
[5] Linde, A. D. 1982 *Phys. Lett.* B **116**, 335, 340.
[6] Albrecht, A. and Steinhardt, P. J. 1982 *Phys. Rev. Lett.* **48**, 1220.
[7] Linde, A. D. 1990 *Particle physics and inflationary cosmology*. Harwood Academic Publishers, London.
[8] Linde, A. D. 2002 Inflationary cosmology and creation of matter in the Universe. In *Modern cosmology* (ed. S. Bonometto, V. Gorini and U. Moschella). Institute of Physics Publishing, Bristol.
[9] Dimopoulos, K. and Valle, J. W. F. 2002 *Astroparticle Phys.* **18**, 287.
[10] Steinhardt, P. J. and Turok, N. 2002 *Science* **296**, 1496, and further references therein.
[11] Peebles, P. J. E. 1993 *Principles of physical cosmology*. Princeton University Press, Princeton, NJ.
[12] Raine, D. J. and Thomas, E. G. 2001 *An introduction to the science of cosmology*. Institute of Physics Publishing, Bristol.

8

Cosmic Microwave Background

In this chapter we shall meet several important observational discoveries. The cosmic microwave background (CMB), which is a consequence of the hot Big Bang and the following radiation-dominated epoch, was discovered in 1964. We discuss this discovery in Section 8.1.

The hot Big Bang also predicts that the CMB radiation should have a blackbody spectrum. Inflation predicts that the mean temperature of the CMB should exhibit minute perturbations across the sky. These predictions were verified by a 1990 satellite experiment, the Cosmic Background Explorer (COBE). Many experiments have since then extended the range of observations with improved precision at different angular scales, most recently the satellite Wilkinson Microwave Anisotropy Probe (WMAP) (see cover picture), whose measurements will be discussed in the rest of this chapter. In Section 8.2 we shall discuss the method of analysing the temperature perturbations.

The temperature perturbations are expected to be associated with even smaller polarization variations, due to Thomson scattering at the LSS. These were first observed by the ground-based Degree Angular Scale Interferometer (DASI) in late 2002 and by WMAP in early 2003. We discuss this in Section 8.3.

The CMB contains a wealth of information about the dynamical parameters of the Universe and on specific features of the theoretical models: general relativity, the standard FLRW cosmology versus other cosmologies, all versions of inflation and its alternatives, dark energy, etc. In Section 8.4 we establish the parameters,

Introduction to Cosmology Third Edition by Matts Roos
© 2003 John Wiley & Sons, Ltd ISBN 0 470 84909 6 (cased) ISBN 0 470 84910 X (pbk)

how they are related to each other, what observational values they have and what information they give about possible cosmological models.

8.1 The CMB Temperature

Predictions. In 1948, *Georg Gamow* (1904–1968), *Ralph Alpher* and *Robert Herman* calculated the temperature at that time of the primordial blackbody radiation which started free streaming at the LSS. They found that the CMB should still exist today, but that it would have cooled in the process of expansion to the very low temperature of $T_0 \approx 5$ K. This corresponds to a photon wavelength of

$$\lambda = \frac{hc}{kT_0} = 2.9 \times 10^{-3} \text{ m}. \tag{8.1}$$

This is in the microwave range of radio waves (see Table A.3). (The term 'microwave' is actually a misnomer, since it does not refer to micrometre wavelengths, but rather to centimetres.)

We can now redo their calculation, using some hindsight. Let us first recall from Equations (4.39) and (4.40) that the expansion rate changed at the moment when radiation and matter contributed equally to the energy density. For our calculation we need to know this equality time, t_{eq}, and the temperature T_{eq}. The radiation energy density is given by Equation (5.47):

$$\varepsilon_{\text{r}} = (g_\gamma + 3g_\nu)\tfrac{1}{2}aT_{\text{eq}}^4. \tag{8.2}$$

The energy density of matter at time T_{eq} is given by Equation (5.26), except that the electron (e^- and e^+) energy needs to be averaged over the spectrum (5.42). We could in principle solve for T_{eq} by equating the radiation and matter densities,

$$\varepsilon_{\text{r}}(T_{\text{eq}}) = \rho_{\text{m}}(T_{\text{eq}}). \tag{8.3}$$

We shall defer solving this to Section 8.4. The temperature T_{eq} corresponds to the crossing of the two lines in the log–log plot in Figure 5.1.

The transition epoch happens to be close to the recombination time (z_{rec} in redshift, see Equation (5.76)) and the LSS (z_{LSS} in redshift, see Equation (5.77)). With their values for t_{eq}, T_{eq} and t_0 and Equation (4.39), Gamow, Alpher and Herman obtained a prediction for the present temperature of the CMB:

$$T_0 = T_{\text{eq}}\left(\frac{t_{\text{eq}}}{t_0}\right)^{2/3} = 2.45 \text{ K}. \tag{8.4}$$

This is very close to the present-day observed value, as we shall see.

Discovery. Nobody paid much attention to the prediction of Gamov *et al.*, because the Big Bang theory was generally considered wildly speculative, and detection of the predicted radiation was far beyond the technical capabilities existing at that time. In particular, their prediction was not known to *Arno Penzias* and

Robert Wilson who, in 1964, were testing a sensitive antenna intended for satellite communication. They wanted to calibrate it in an environment free of all radiation, so they chose a wavelength of $\lambda = 0.0735$ m in the relatively quiet window between the known emission from the Galaxy at longer wavelengths and the emission at shorter wavelengths from the Earth's atmosphere. They also directed the antenna high above the galactic plane, where scattered radiation from the Galaxy would be minimal.

To their consternation and annoyance they found a constant low level of background noise in every direction. This radiation did not seem to originate from distant galaxies, because in that case they would have seen an intensity peak in the direction of the nearby M31 galaxy in Andromeda. It could also not have originated in Earth's atmosphere, because such an effect would have varied with the altitude above the horizon as a function of the thickness of the atmosphere.

Thus Penzias and Wilson suspected technical problems with the antenna (in which a couple of pigeons turned out to be roosting) or with the electronics. All searches failing, they finally concluded, correctly, that the Universe was uniformly filled with an 'excess' radiation corresponding to a blackbody temperature of 3.5 K, and that this radiation was isotropic and unpolarized within their measurement precision.

At Princeton University, a group of physicists led by *Robert Dicke* (1916–1997) had at that time independently arrived at the conclusion of Gamow and collaborators, and they were preparing to measure the CMB radiation when they heard of the remarkable 3.5 K 'excess' radiation. The results of Penzias and Wilson's measurements were immediately understood and they were subsequently published (in 1965) jointly with an article by Dicke and collaborators which explained the cosmological implications. The full story is told by Peebles [1], who was a student of Dicke at that time. Penzias and Wilson (but not Gamow or Dicke) were subsequently awarded the Nobel prize in 1978 for this discovery.

This evidence for the 15-Gyr-old echo of the Big Bang counts as the most important discovery in cosmology since Hubble's law. In contrast to all radiation from astronomical bodies, which is generally hotter, and which has been emitted much later, the CMB has existed since the era of radiation domination. It is hard to understand how the CMB could have arisen without the cosmic material having once been highly compressed and exceedingly hot. There is no known mechanism at any time after decoupling that could have produced a blackbody spectrum in the microwave range, because the Universe is transparent to radio waves.

Spectrum. In principle, one intensity measurement at an arbitrary wavelength of the blackbody spectrum (5.3) is sufficient to determine its temperature, T, because this is the only free parameter. On the other hand, one needs measurements at different wavelengths to establish that the spectrum is indeed blackbody.

It is easy to see that a spectrum which was blackbody at time t with temperature T will still be blackbody at time t' when the temperature has scaled to

$$T' = T\frac{a(t)}{a(t')}. \tag{8.5}$$

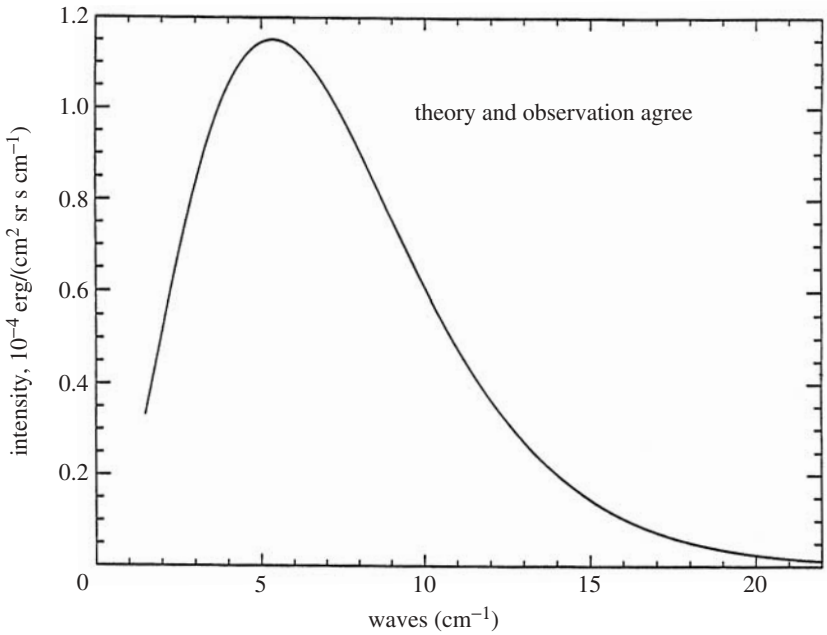


Figure 8.1 Spectrum of the CMB from data taken with the FIRAS instrument aboard NASA's COBE. Reproduced from reference [3] by permission of the COBE Science Working Group.

This is so because, in the absence of creation or annihilation processes, the number of photons, $n_\gamma a^3(t)$, is conserved. Thus the number density $\mathrm{d}n_\gamma(\nu)$ in the frequency interval $(\nu, \nu + \mathrm{d}\nu)$ at time t transforms into the number density at time t',

$$\mathrm{d}n'_\gamma(\nu') = \left(\frac{a(t)}{a(t')}\right)^3 \mathrm{d}n_\gamma(\nu). \tag{8.6}$$

Making use of Equations (5.3) and (8.5), the distribution at time t' becomes

$$\mathrm{d}n'_\gamma(\nu') = \frac{8\pi}{3c^3} \frac{\mathrm{d}\nu'^3}{\mathrm{e}^{h\nu'/kT'} - 1}, \tag{8.7}$$

which is precisely the blackbody spectrum at temperature T'.

Although several accurate experiments since Penzias and Wilson have confirmed the temperature to be near 3 K by measurements at different wavelengths, the conclusive demonstration that the spectrum is indeed also blackbody in the region masked by radiation from the Earth's atmosphere was first made by a dedicated instrument, the Far Infrared Absolute Spectrophotometer (FIRAS) aboard the COBE satellite launched in 1989 [2]. The present temperature, T_0, is taken to be

$$T_0 = 2.725 \pm 0.001 \text{ K}. \tag{8.8}$$

The spectrum reported by the COBE team in 1993 [3], shown in Figure 8.1, matches the predictions of the hot Big Bang theory to an extraordinary degree. The measurement errors on each of the 34 wavelength positions are so small that

they cannot be distinguished from the theoretical blackbody curve. It is worth noting that such a pure blackbody spectrum had never been observed in laboratory experiments. All theories that attempt to explain the origin of large-scale structure seen in the Universe today must conform to the constraints imposed by these COBE measurements.

The vertical scale in Figure 8.1 gives the *intensity* $I(1/\lambda)$ of the radiation, that is, the power per unit inverse wavelength interval arriving per unit area at the observer from one steradian of sky. In SI units this is 10^{-9} J m^{-1} sr^{-1} s^{-1}. This quantity is equivalent to the intensity per unit frequency interval, $I(\nu)$. One can transform from dλ to dν by noting that $I(\nu)\,\mathrm{d}\nu = I(\lambda)\,\mathrm{d}\lambda$, from which

$$I(\lambda) = \frac{\nu^2}{c} I(\nu). \tag{8.9}$$

The relation between energy density ε_r and total intensity, integrated over the spectrum, is

$$\varepsilon_\mathrm{r} = \frac{4\pi}{c} \int I(\nu)\,\mathrm{d}\nu. \tag{8.10}$$

Energy and Entropy Density. Given COBE's precise value of T_0, one can determine several important quantities. From Equation (5.47) one can calculate the present energy density of radiation

$$\varepsilon_{r,0} = \tfrac{1}{2} g_* a_\mathrm{S} T_0^4 = 2.604 \times 10^5 \text{ eV m}^{-3}. \tag{8.11}$$

The corresponding density parameter then has the value

$$\Omega_\mathrm{r} = \frac{\varepsilon_{r,0}}{\rho_\mathrm{c}} = 2.471 \times 10^{-5}\ \mathrm{h}^{-2}, \tag{8.12}$$

using the value of ρ_c from Equation (1.31). Obviously, the radiation energy is very small today and far from the value $\Omega_0 = 1$ required to close the Universe.

The present value of the entropy density is

$$s = \frac{4}{3}\frac{\varepsilon_{r,0}}{kT} = \frac{4}{3}\frac{g_{*\mathrm{S}}}{2}\frac{a_\mathrm{S} T^4}{kT} = 2.890 \times 10^9 \text{ m}^{-3}. \tag{8.13}$$

Recall (from the text immediately after Equation (5.73)) that the (T_ν/T) dependence of $g_{*\mathrm{S}}$ is a power of three rather than a power of four, so the factor $(\frac{4}{11})^{4/3}$ becomes just $\frac{4}{11}$ and $g_{*\mathrm{S}}$ becomes 3.91.

The present number density of CMB photons is given directly by Equation (5.5):

$$N_\gamma = \zeta(3)\frac{2}{\pi^2}\left(\frac{kT}{c\hbar}\right)^3 = 4.11 \times 10^8 \text{ photons m}^{-3}. \tag{8.14}$$

Neutrino Number Density. Now that we know T_0 and N_γ we can obtain the neutrino temperature $T_\nu = 1.949$ K from Equation (5.71) and the neutrino number density per neutrino species from Equation (5.72),

$$N_\nu = \tfrac{3}{11} N_\gamma = 1.12 \times 10^8 \text{ neutrinos m}^{-3}. \tag{8.15}$$

For three species of relic neutrinos with average mass $\langle m_\nu \rangle$, Equation (5.74) can be used to cast the density parameter in the form

$$\Omega_\nu = \frac{3\langle m_\nu \rangle}{94.0h^2 \text{ eV}}. \tag{8.16}$$

8.2 Temperature Anisotropies

The Dipole Anisotropy. The temperature measurement of Penzias and Wilson's antenna was not very precise by today's standards. Their conclusion about the isotropy of the CMB was based on an accuracy of only 1.0 K. When the measurements improved over the years it was found that the CMB exhibited a *dipole anisotropy*. The temperature varies minutely over the sky in such a way that it is maximally blueshifted in one direction (call it α) and maximally redshifted in the opposite direction ($\alpha + 180°$). In a direction $\alpha + \theta$ it is

$$T(\theta) = T(\alpha)(1 + v \cos\theta), \tag{8.17}$$

where v is the amplitude of the dipole anisotropy. Although this shift is small, only $vT(\alpha) \approx 3.35$ mK, it was measured with an accuracy better than 1% by the Differential Microwave Radiometer (DMR) instrument on board the COBE satellite [4].

At the end of Chapter 5 we concluded that the hot Big Bang cosmology predicted that the CMB should be essentially isotropic, since it originated in the LSS, which has now receded to a redshift of $z \simeq 1100$ in all directions. Note that the most distant astronomical objects have redshifts up to $z = 7$. Their distance in time to the LSS is actually much closer than their distance to us.

In the standard model the expansion is spherically symmetric, so it is quite clear that the dipole anisotropy cannot be of cosmological origin. Rather, it is well explained by our motion 'against' the radiation in the direction of maximal blueshift with relative velocity v.

Thus there is a frame in which the CMB is isotropic—not a rest frame, since radiation cannot be at rest. This frame is then comoving with the expansion of the Universe. We referred to it in Section 2.2, where we noted that, to a fundamental observer at rest in the comoving frame, the Universe must appear isotropic if it is homogeneous. Although general relativity was constructed to be explicitly frame independent, the comoving frame in which the CMB is isotropic is observationally convenient. The fundamental observer is at position B in Figure 8.2.

The interpretation today is that not only does the Earth move around the Sun, and the Solar System participates in the rotation of the Galaxy, but also the Galaxy moves relative to our Local Galaxy Group, which in turn is falling towards a centre behind the Hydra–Centaurus supercluster in the constellation Virgo. From the observation that our motion relative to the CMB is about 365 km s^{-1}, these velocity vectors add up to a peculiar motion of the Galaxy of about 550 km s^{-1}, and a peculiar motion of the Local Group of about 630 km s^{-1} [5]. Thus the dipole anisotropy seen by the Earth-based observer A in Figure 8.2 tells us that we and the Local Group are part of a larger, gravitationally bound system.

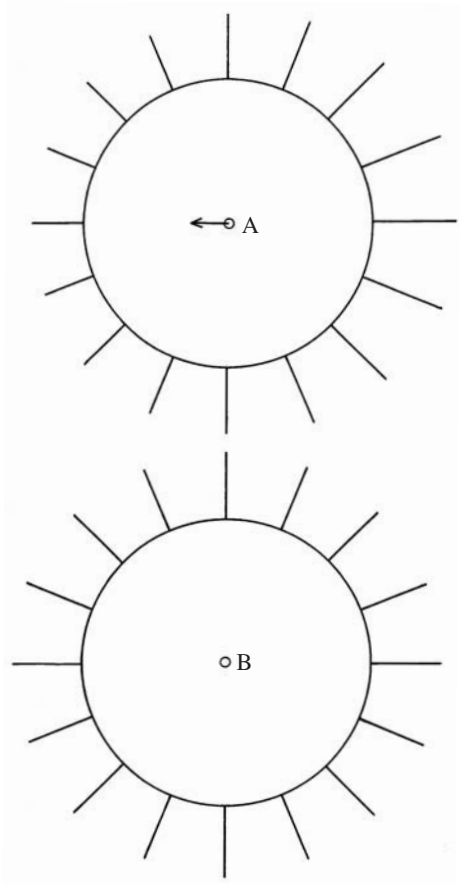


Figure 8.2 The observer A in the solar rest frame sees the CMB to have dipole anisotropy—the length of the radial lines illustrate the CMB intensity—because he is moving in the direction of the arrow. The fundamental observer at position B has removed the anisotropy.

Multipole Analysis. Temperature fluctuations around a mean temperature T_0 in a direction α on the sky can be analysed in terms of the *autocorrelation function* $C(\theta)$, which measures the product of temperatures in two directions $\boldsymbol{m}, \boldsymbol{n}$ separated by an angle θ and averaged over all directions α,

$$C(\theta) = \left\langle \frac{\delta T(\boldsymbol{m})}{T_0} \frac{\delta T(\boldsymbol{n})}{T_0} \right\rangle, \qquad \boldsymbol{m} \cdot \boldsymbol{n} = \cos\theta. \tag{8.18}$$

For small angles (θ) the temperature autocorrelation function can be expressed as a sum of *Legendre polynomials* $P_\ell(\theta)$ of order ℓ, the *wavenumber*, with coefficients or *powers* a_ℓ^2,

$$C(\theta) = \frac{1}{4\pi} \sum_{\ell=2}^{\infty} a_\ell^2 (2\ell + 1) P_\ell(\cos\theta). \tag{8.19}$$

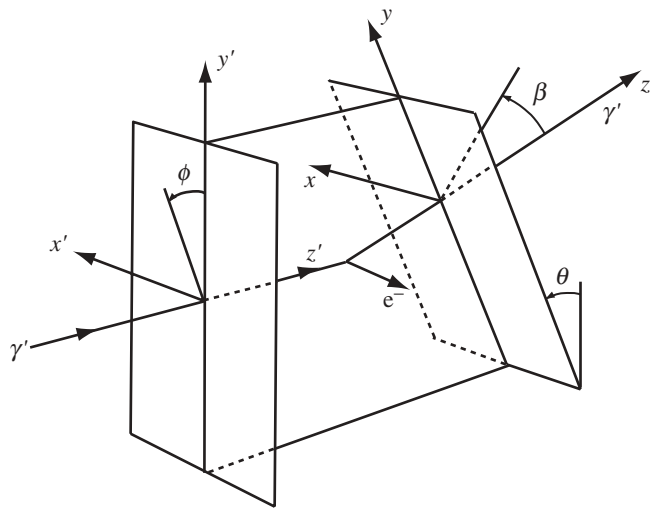


Figure 8.3 The geometry used in the text for describing the polarization of an incoming unpolarized plane wave photon, γ' in the (x', y')-plane, which is Thomson scattering against an electron, and subsequently propagating as a polarized plane wave photon, γ, in the z-direction.

All analyses start with the quadrupole mode $\ell = 2$ because the $\ell = 0$ monopole mode is just the mean temperature over the observed part of the sky, and the $\ell = 1$ mode is the dipole anisotropy. Higher multipoles correspond to fluctuations on angular scales

$$\theta \simeq \frac{60^\circ}{\ell}.$$

In the analysis, the powers a_ℓ^2 are adjusted to give a best fit of $C(\theta)$ to the observed temperature. The resulting distribution of a_ℓ^2 values versus ℓ is called the *power spectrum* of the fluctuations. The higher the angular resolution, the more terms of high ℓ must be included. Anisotropies on the largest angular scales corresponding to quadrupoles are manifestations of truly primordial gravitational structures.

For the analysis of temperature perturbations over large angles, the Legendre polynomial expansion (8.19) will not do; one has to use tensor spherical harmonics. Consider a plane wave of unpolarized photons with electric field vector $\boldsymbol{E}$ propagating in the z direction (cf. Figure 8.3). The components of $\boldsymbol{E}$ can be taken from Equation (5.7), and the intensity $\boldsymbol{I}$ of the radiation field from Equation (5.8). Generalizing from a plane wave to a radiation field $\boldsymbol{E}(\boldsymbol{n})$ in the direction $\boldsymbol{n}$, the temperature $T(\boldsymbol{n})$ can be expanded in spherical harmonics

$$T(\boldsymbol{n}) = T_0 + \sum_{\ell=1}^{\infty} \sum_{m=-\ell}^{\ell} a_{\ell m}^{\mathrm{T}} Y_{\ell m}(\boldsymbol{n}), \tag{8.20}$$

where $a_{\ell m}^{\mathrm{T}}$ are the powers or *temperature multipole components*. These can be determined from the observed temperature $T(\boldsymbol{n})$ using the orthonormality prop-

erties of the spherical harmonics,

$$a^{\mathrm{T}}_{\ell m} = \frac{1}{T_0}\int \mathrm{d}\boldsymbol{n}\, T(\boldsymbol{n}) Y^*_{\ell m}(\boldsymbol{n}). \tag{8.21}$$

Expressing the autocorrelation function C as a power spectrum in terms of the multipole components, the average of all statistical realizations of the distribution becomes

$$\langle a^{\mathrm{T}*}_{\ell m} a^{\mathrm{T}}_{\ell' m'}\rangle = C^{\mathrm{T}}_{\ell}\delta_{\ell\ell'}\delta_{mm'} = C^{\mathrm{T}}_{\ell}. \tag{8.22}$$

The last step follows from statistical isotropy which requires

$$\delta_{\ell\ell'} = 1 \quad \text{and} \quad \delta_{mm'} = 1.$$

Sources of Anisotropies. Let us now follow the fate of the scalar density perturbations generated during inflation, which subsequently froze and disappeared outside the (relatively slowly expanding) horizon. For wavelengths exceeding the horizon, the distinction between curvature (adiabatic) and isocurvature (isothermal) perturbations is important. Curvature perturbations are true energy density fluctuations or fluctuations in the local value of the spatial curvature. These can be produced, for example, by the quantum fluctuations that are blown up by inflation. By the equivalence principle all components of the energy density (matter, radiation) are affected. Isocurvature fluctuations, on the other hand, are not true fluctuations in the energy density but are characterized by fluctuations in the form of the local equation of state, for example, spatial fluctuations in the number of some particle species. These can be produced, for example, by cosmic strings that perturb the local equation of state. As long as an isocurvature mode is super-horizon, physical processes cannot re-distribute the energy density.

When the Universe arrived at the radiation- and matter-dominated epochs, the Hubble expansion of the horizon reveals these perturbations. Once inside the horizon, the crests and troughs can again communicate, setting up a pattern of standing acoustic waves in the baryon–photon fluid. The tight coupling between radiation and matter density causes the adiabatic perturbations to oscillate in phase. After decoupling, the perturbations in the radiation field no longer oscillate, and the remaining standing acoustic waves are visible today as perturbations to the mean CMB temperature at degree angular scales.

Curvature and isocurvature fluctuations behave differently when they are super-horizon: isocurvature perturbations cannot grow, while curvature perturbations can. Once an isocurvature mode passes within the horizon, however, local pressure can move energy density and can convert an isocurvature fluctuation into a true energy-density perturbation. For sub-horizon modes the distinction becomes unimportant and the Newtonian analysis applies to both. However, isocurvature fluctuations do not lead to the observed acoustic oscillations seen in Figure 8.4 (they do not peak in the right place), whereas the adiabatic picture is well confirmed.

At the LSS, crests in the matter density waves imply higher gravitational potential. As we learned in Section 2.5, photons ‘climbing out’ of overdense regions will

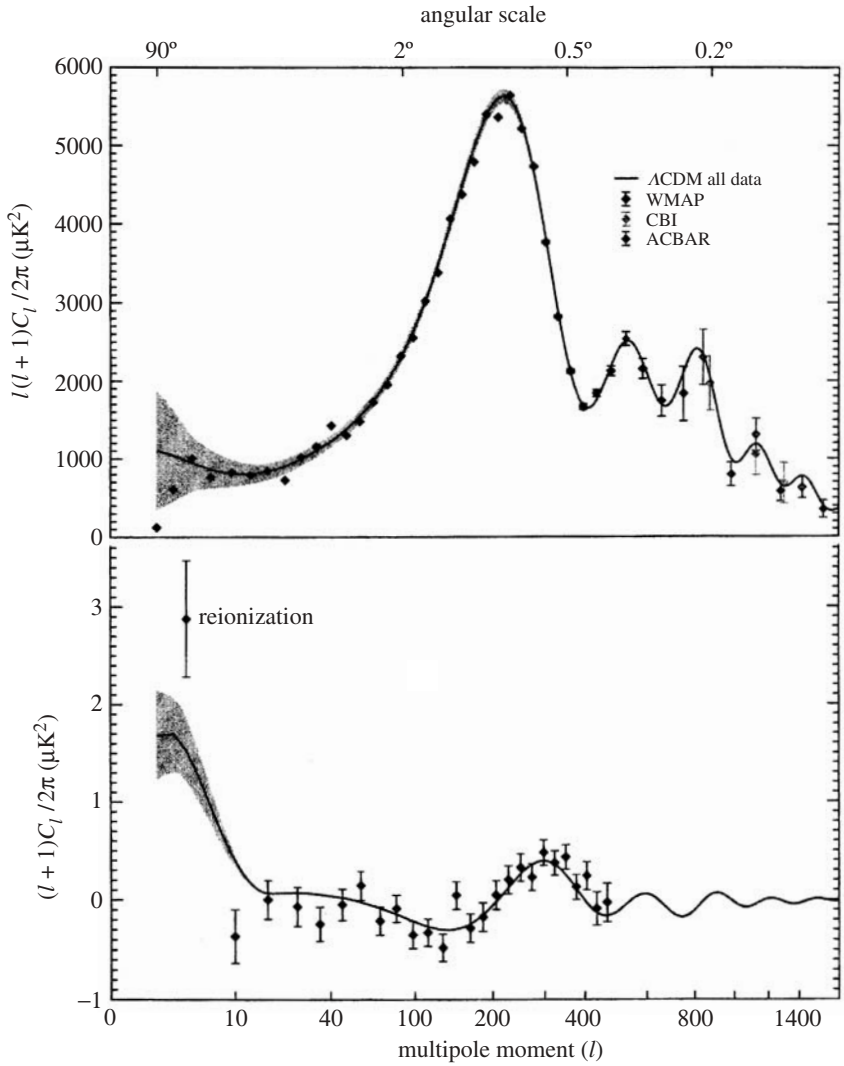


Figure 8.4 The best-fit power spectra of CMB fluctuations as a function of angular scale (top) and wavenumber (bottom). The upper figure shows the temperature (T) power spectrum, and the lower figure the temperature–polarization (TE) cross-power spectrum. Note that the latter is not multiplied by the additional factor ℓ . The grey shading represents the 1σ cosmic variance. For further details, see [6]. Reproduced from reference [6] by permission of the WMAP Team.

be redshifted by an amount given by Equation (2.79), but this is partially offset by the higher radiation temperature in them. This source of anisotropy is called the *Sachs–Wolfe effect*. Inversely, photons emitted from regions of low density 'roll down' from the gravitational potential, and are blueshifted. In the long passage to us they may traverse further regions of gravitational fluctuations, but then their frequency shift upon entering the potential is compensated for by an opposite

frequency shift when leaving it (unless the Hubble expansion causes the potential to change during the traverse). They also suffer a time dilation, so one effectively sees them at a different time than unshifted photons. Thus the CMB photons preserve a 'memory' of the density fluctuations at emission, manifested today as temperature variations at large angular scales. An anisotropy of the CMB of the order of $\delta T/T \approx 10^{-5}$ is, by the Sachs–Wolfe effect, related to a mass perturbation of the order of $\delta \approx 10^{-4}$ when averaged within one Hubble radius.

The gravitational redshift and the time dilation both contribute to $\delta T/T_0$ by amounts which are linearly dependent on the density fluctuations $\delta\rho/\rho$, so the net effect is given by

$$\frac{\delta T}{T} \simeq \frac{1}{3}\left(\frac{L_{\rm dec}}{ct_{\rm dec}}\right)^2 \frac{\delta\rho}{\rho}, \tag{8.23}$$

where $L_{\rm dec}$ is the size of the structure at decoupling time $t_{\rm dec}$ (corresponding to $z_{\rm dec}$ in Equation (5.78)). (Note that Equation (8.23) is strictly true only for a critical universe with zero cosmological constant.)

The space-time today may also be influenced by primordial fluctuations in the metric tensor. These would have propagated as gravitational waves, causing anisotropies in the microwave background and affecting the large-scale structures in the Universe. High-resolution measurements of the large-angle microwave anisotropy are expected to be able to resolve the tensor component from the scalar component and thereby shed light on our inflationary past.

Further sources of anisotropies may be due to variations in the values of cosmological parameters, such as the cosmological constant, the form of the quintessence potential, and local variations in the time of occurrence of the LSS.

Discovery. For many years microwave experiments tried to detect temperature variations on angular scales ranging from a few arc minutes to tens of degrees. Ever increasing sensitivities had brought down the limits on $\delta T/T$ to near 10^{-5} without finding any evidence for anisotropy until 1992. At that time, the first COBE observations of large-scale CMB anisotropies bore witness of the spatial distribution of inhomogeneities in the Universe on comoving scales ranging from a few hundred Mpc up to the present horizon size, without the complications of cosmologically recent evolution. This is inaccessible to any other astronomical observations.

On board the COBE satellite there were several instruments, of which one, the DMR, received at three frequencies and had two antennas with 7° opening angles directed 60° apart. This instrument compared the signals from the two antennas, and it was sensitive to anisotropies on large angular scales, corresponding to multipoles $\ell < 30$. Later radio telescopes were sensitive to higher multipoles, so one now has a detailed knowledge of the multipole spectrum up to $\ell = 2800$.

The most precise recent results are shown in the upper part of Figure 8.4. At low ℓ, the temperature–power spectrum is smooth, caused by the Sachs–Wolfe effect. Near $\ell = 200$ it rises towards the first and dominant peak of a series of *Sakharov oscillations*, also confusingly called the *Doppler peak*. They are basically caused by

density perturbations which oscillate as acoustic standing waves inside the LSS horizon. The exact form of the power spectrum is very dependent on assumptions about the matter content of the Universe; thus careful measurement of its shape yields precise information about many dynamical parameters. For details on the results included in the figure, see reference [6].

The definitive DMR results [4] cover four years of measurements of eight complete mappings of the full sky followed by the above spherical harmonic analysis. The CMB anisotropies found correspond to temperature variations of

$$\delta T = 29 \pm 1\ \mu\text{K}, \quad \text{or } \delta T/T = 1.06 \times 10^{-5}. \tag{8.24}$$

Half of the above temperature variations, or $\delta T = 15.3\ \mu$K, could be ascribed to quadrupole anisotropy at the 90° angular scale. Although some quadrupole anisotropy is kinetic, related to the dipole anisotropy and the motion of Earth, this term could be subtracted. The remainder is then quadrupole anisotropy of purely cosmological origin.

Since the precision of the COBE measurements surpassed all previous experiments one can well understand that such small temperature variations had not been seen before. The importance of this discovery was succinctly emphasized by the COBE team who wrote that 'a new branch of astronomy has commenced'. The story of the COBE discoveries have been fascinatingly narrated by George Smoot [7].

8.3 Polarization Anisotropies

Thomson Scattering. Elastic scattering of photons from free electrons, Equation (5.30) is called *Thomson scattering* or Compton scattering, the latter being used for higher frequencies. In Section 5.5 we ignored the fate of the primordial photons, noting that they were thermalized by this process before decoupling. We also noted that unpolarized photons are polarized by the anisotropic Thomson scattering process, but as long as the photons continue to meet free electrons their polarization is washed out, and no net polarization is produced. At a photon's last scattering, however, the induced polarization remains and the subsequently free-streaming photon possesses a quadrupole moment ($\ell = 2$).

The perturbations to the baryon density and the radiation temperature in the tightly coupled baryon–photon fluid are scalar, thus corresponding to a monopole moment ($\ell = 0$). As we saw in the previous section, the radiation field also exhibits dipole perturbations ($\ell = 1$) which are coupled to the baryon bulk velocity, but there are no vector or tensor perturbations. Tensor perturbations would be due to gravitational waves, which have not been observed with present-day detectors.

The quadrupole moment possessed by free-streaming photons couples more strongly to the bulk velocity (the peculiar velocities) of the baryon–photon fluid than to the density. Therefore, the photon density fluctuations generate temperature fluctuations, while the velocity gradient generates polarization fluctuations.

Let us now make use of the description of photons in Section 5.1 and the Stokes parameters (5.8). The parameter I, which describes the intensity of radiation, is,

like V, a physical observable independent of the coordinate system. In contrast, the parameters Q and U depend on the orientation of the coordinate system. In the geometry of Figure 8.3, the coordinates x', y' define a plane wave of incoming radiation propagating in the z' direction (primes are used for unscattered quantities). The incoming photon y' then Thomson scatters against an electron and the outgoing photon y continues as a plane wave in a new direction, z.

It follows from the definition of the Stokes parameters Q and U that a rotation of the x'- and y'-axes in the incoming plane by the angle ϕ transforms them into

$$Q(\phi) = Q\cos(2\phi) + U\sin(2\phi), \qquad U(\phi) = -Q\sin(2\phi) + U\cos(2\phi). \quad (8.25)$$

We left it as an exercise (Chapter 5, Problem 2) to demonstrate that $Q^2 + U^2$ is invariant under the rotation (8.25). It follows from this invariance that the polarization P is a second rank tensor of the form

$$P = \frac{1}{2}\begin{pmatrix} Q & U - \mathrm{i}V \\ U + \mathrm{i}V & -Q \end{pmatrix}. \quad (8.26)$$

Thus the polarization is not a vector quantity with a direction unlike the electric field vector $\boldsymbol{E}$.

Let us now see how Thomson scattering of the incoming, unpolarized radiation generates linear polarization in the (x, y)-plane of the scattered radiation (we follow closely the pedagogical review of A. Kosowsky [8]). The differential scattering cross-section, defined as the radiated intensity I divided by the incoming intensity I' per unit solid angle Ω and cross-sectional area σ_{B}, is given by

$$\frac{\mathrm{d}\sigma_{\mathrm{T}}}{\mathrm{d}\Omega} = \frac{I}{I'} = \frac{3\sigma_{\mathrm{T}}}{8\pi\sigma_{\mathrm{B}}}|\boldsymbol{i}' \cdot \boldsymbol{i}|^2 \equiv K|\boldsymbol{i}' \cdot \boldsymbol{i}|^2. \quad (8.27)$$

Here σ_{T} is the total Thomson cross-section, the vectors $\boldsymbol{i}'$, $\boldsymbol{i}$ are the unit vectors in the incoming and scattered radiation planes, respectively (cf. Figure 8.3), and we have lumped all the constants into one constant, K. The Stokes parameters of the outgoing radiation then depend solely on the nonvanishing incoming parameter I',

$$I = KI'(1 + \cos^2\theta), \qquad Q = KI'\sin^2\theta, \qquad U = 0, \quad (8.28)$$

where θ is the scattering angle. By symmetry, Thomson scattering can generate no circular polarization, so $V = 0$ always.

The net polarization produced in the direction $\boldsymbol{z}$ from an incoming field of intensity $I'(\theta, \phi)$ is determined by integrating Equations (8.28) over all incoming directions. Note that the coordinates for each incoming direction must be rotated some angle ϕ about the z-axis as in Equations (8.25), so that the outgoing Stokes parameters all refer to a common coordinate system. The result is then [8]

$$I(\boldsymbol{z}) = \tfrac{1}{2}K\int \mathrm{d}\Omega(1 + \cos^2\theta)I'(\theta, \phi), \quad (8.29)$$

$$Q(\boldsymbol{z}) - \mathrm{i}U(\boldsymbol{z}) = \tfrac{1}{2}K\int \mathrm{d}\Omega\ \sin^2\theta \mathrm{e}^{2\mathrm{i}\phi}I'(\theta, \phi). \quad (8.30)$$

Expanding the incident radiation intensity in spherical coordinates,

$$I'(\theta,\phi) = \sum_{\ell m} a_{\ell m} Y_{\ell m}(\theta,\phi), \tag{8.31}$$

leads to the following expressions for the outgoing Stokes parameters:

$$I(\boldsymbol{z}) = \tfrac{1}{2}K\left(\tfrac{8}{3}\sqrt{\pi}a_{00} + \tfrac{4}{3}\sqrt{\tfrac{1}{5}\pi}a_{20}\right), \tag{8.32}$$

$$Q(\boldsymbol{z}) - \mathrm{i}U(\boldsymbol{z}) = 2K\sqrt{\frac{2\pi}{15}}a_{22}. \tag{8.33}$$

Thus, if there is a nonzero quadrupole moment a_{22} in the incoming, unpolarized radiation field, it will generate linear polarization in the scattering plane. To determine the outgoing polarization in some other scattering direction, $\boldsymbol{n}$, making an angle β with $\boldsymbol{z}$, one expands the incoming field in a coordinate system rotated through β. This derivation requires too much technical detail to be carried out here, so we only state the result [8]:

$$Q(\boldsymbol{n}) - \mathrm{i}U(\boldsymbol{z}\boldsymbol{n}) = K\sqrt{\tfrac{1}{5}\pi}a_{20}\sin^2\beta. \tag{8.34}$$

Multipole Analysis. The tensor harmonic expansion (8.20) for the radiation temperature T and the temperature multipole components $a^{\mathrm{T}}_{(\ell m)}$ in (8.21) can now be completed with the corresponding expressions for the polarization tensor P. From the expression (8.26) its components are

$$\begin{aligned} P_{ab}(\boldsymbol{n}) &= \frac{1}{2}\begin{pmatrix} Q(\boldsymbol{n}) & -U(\boldsymbol{n})\sin\theta \\ -U(\boldsymbol{n})\sin\theta & -Q(\boldsymbol{n})\sin^2\theta \end{pmatrix} \\ &= T_0 \sum_{\ell=2}^{\infty}\sum_{m=-\ell}^{\ell}\left[a^{\mathrm{E}}_{(\ell m)}Y^{\mathrm{E}}_{(\ell m)ab}(\boldsymbol{n}) + a^{\mathrm{B}}_{(\ell m)}Y^{\mathrm{B}}_{(\ell m)ab}(\boldsymbol{n})\right]. \end{aligned} \tag{8.35}$$

The existence of the two modes (superscripted) E and B is due to the fact that the symmetric traceless tensor (8.35) describing linear polarization is specified by two independent Stokes parameters, Q and U. This situation bears analogy with the electromagnetic vector field, which can be decomposed into the gradient of a scalar field (E for electric) and the curl of a vector field (B for magnetic). The source of the E-modes is Thomson scattering. The sources of the B-modes are gravitational waves entailing tensor perturbations, and E-modes which have been deformed by gravitational lensing of large-scale structures in the Universe.

In analogy with Equation (8.21), the polarization multipole components are

$$a^{\mathrm{E}}_{\ell m} = \frac{1}{T_0}\int \mathrm{d}\boldsymbol{n}\, P_{ab}(\boldsymbol{n})Y^{\mathrm{E}\,ab*}_{\ell m}(\boldsymbol{n}), \tag{8.36}$$

$$a^{\mathrm{B}}_{\ell m} = \frac{1}{T_0}\int \mathrm{d}\boldsymbol{n}\, P_{ab}(\boldsymbol{n})Y^{\mathrm{B}\,ab*}_{\ell m}(\boldsymbol{n}). \tag{8.37}$$

The three sets of multipole moments $a^{\mathrm{T}}_{\ell m}$, $a^{\mathrm{E}}_{\ell m}$ and $a^{\mathrm{B}}_{\ell m}$ fully describe the temperature and polarization map of the sky; thus, they are physical observables.

There are then six power spectra in terms of the multipole components: the temperature spectrum in Equation (8.22) and five spectra involving linear polarization. The full set of physical observables is then

$$\left.\begin{aligned} C_\ell^{\mathrm{T}} &= \left\langle a_{\ell m}^{\mathrm{T}*} a_{\ell' m'}^{\mathrm{T}} \right\rangle, \qquad C_\ell^{\mathrm{E}} = \left\langle a_{\ell m}^{\mathrm{E}*} a_{\ell' m'}^{\mathrm{E}} \right\rangle, \qquad C_\ell^{\mathrm{B}} = \left\langle a_{\ell m}^{\mathrm{B}*} a_{\ell' m'}^{\mathrm{B}} \right\rangle, \\ C_\ell^{\mathrm{TE}} &= \left\langle a_{\ell m}^{\mathrm{T}*} a_{\ell' m'}^{\mathrm{E}} \right\rangle, \qquad C_\ell^{\mathrm{TB}} = \left\langle a_{\ell m}^{\mathrm{T}*} a_{\ell' m'}^{\mathrm{B}} \right\rangle, \qquad C_\ell^{\mathrm{EB}} = \left\langle a_{\ell m}^{\mathrm{E}*}\, a_{\ell' m'}^{\mathrm{B}} \right\rangle. \end{aligned}\right\} \tag{8.38}$$

For further details on polarization, see Kosowsky [8].

Observations. Thus polarization delivers six times more information than temperature alone. In practice, at most four of the power spectra will become available, since the intensity of polarization is so much weaker than that of temperature, and the components C_ℓ^{TB} and C_ℓ^{EB} are then very difficult to observe.

The first observations of the polarization spectra C_ℓ^{E} and C_ℓ^{TE} were made by the South Pole-based Degree Angular Scale Interferometer DASI [9] and the WMAP satellite [6]. In the lower part of Figure 8.4 we plot the temperature–polarization (TE) cross-power spectrum from the first year of WMAP observations.

8.4 Model Testing and Parameter Estimation. II

The parameters required to model CMB are some subset of the parameters H, $\Omega_{\mathrm{m}}h^2$, $\Omega_{\mathrm{b}}h^2$, Ω_0, the optical depth τ to LSS, the amplitude A of the power spectrum, the scalar and tensor power indices n_{s}, n_{t} in Equation (7.57), the energy variation of the scalar index $\mathrm{d}n_{\mathrm{s}}/\mathrm{d}k$, and the linear theory amplitude of fluctuations σ_8 within 8 Mpc h^{-1} spheres at $z = 0$.

The primordial fluctuations are assumed to be Gaussian random phase, since no evidence to the contrary has been found. Note the qualitative feature in Figure 8.4 that the TE component is zero at the temperature–power spectrum maxima, as is expected (the polarization is maximal at velocity maxima and minimal at temperature maxima, where the velocities are minimal), and that it exhibits a significant large-angle anti-correlation dip at $\ell \approx 150$, a distinctive signature of super-horizon adiabatic fluctuations.

Re-ionization. Since polarization originated in the LSS when the horizon was about 1.12° of our present horizon, the polarization fluctuations should only be visible in multipoles

$$\ell > \frac{60^\circ}{1.12^\circ} \approx 54.$$

But WMAP also observes a strong signal on large angular scales, $\ell < 10$. This can only be due to re-ionization later than LSS, when the radiation on its way to us traversed ionized hydrogen clouds, heated in the process of gravitational contraction. The effect of CMB re-ionization is called the *Sunyaev–Zel'dovich* Effect

(SZE) (*Yakov B. Zel'dovich*, 1914–1987). As a consequence of the SZE, the CMB spectrum is distorted, shifting towards higher energy.

From the size of this spectral shift WMAP estimates a value for the *optical depth* to the effective re-ionization clouds, τ, which is essentially independent of cosmological modelling,

$$\tau = 0.17 \pm 0.04, \tag{8.39}$$

but strongly degenerate with n_s. This corresponds to re-ionization by an early generation of stars at $z_r = 20 \pm 5$. It could still be that this picture is simplistic, the re-ionization may have been a complicated process in a clumpy medium, involving several steps at different redshifts. We shall continue the discussion of this effect in Section 9.2.

Power-Spectrum Parameters. The positions and amplitudes of the peaks and troughs in the temperature (T) and the temperature–polarization (TE) cross-power spectrum in Figure 8.4 contain a wealth of information on cosmological parameters.

The first acoustic T peak determines the scale ℓ of the time when matter compressed for the first time after t_{dec}. The position in ℓ-space is related to the parameters n_s, $\Omega_m h^2$ and $\Omega_b h^2$ (Note that the physical matter density Ω_m includes the baryon density Ω_b.) The amplitude of the first peak is positively correlated to $\Omega_m h^2$ and the amplitude of the second peak is negatively correlated to $\Omega_b h^2$ but, to evaluate Ω_m and Ω_b, one needs to know a value for h, which one can take from Section 1.4. Increasing n_s increases the ratio of the second peak to the first peak. At fixed n_s, however, this ratio determines Ω_b/Ω_m. The amplitudes also put limits on $\Omega_\nu h^2$.

In $(\Omega_m, \Omega_\Lambda)$-space, the CMB data determine $\Omega_0 = \Omega_m + \Omega_\lambda$ most precisely, whereas supernova data (discussed in Section 4.4) determine $\Omega_\Lambda - \Omega_m$ most precisely. Combining both sets of data with data on large-scale structures from 2dFGRS [10] (discussed in Chapter 9) which depend on n_s, $\Omega_m h$, $\Omega_b h$ and which put limits on $\Omega_\nu h$, one breaks the CMB power spectrum parameter degeneracies and improves the precision.

It is too complicated to describe the simultaneous fit to all data and the evaluation of the parameter values here, so we just quote the results published by the WMAP Team [6] with some comments. WMAP includes information also from the CMB detectors Cosmic Background Imager [11] and the Arcminute Cosmology Bolometer Array Receiver [12] and the matter–power spectrum at $z \sim 3$ as measured by the Lyα absorption in intergalactic hydrogen clouds (the *Lyman-α forest*). Note that all errors quoted refer to single-parameter fits at 68% confidence, marginalized over all the other parameters. This results in too small errors if one is interested in two- or many-dimensional confidence regions.

WMAP finds that the scalar index n_s is a slowly varying function of the spectral index k. This slow variation came as a surprise, because inflationary models require n_s to be a constant. WMAP chooses to quote two parameters n_s at

0.05 Mpc^{-1} and its first derivative, which obtain the values

$$n_s(0.05\ \mathrm{Mpc}^{-1}) = 0.93 \pm 0.03, \qquad \frac{\mathrm{d}n_s}{\mathrm{d}k} = -0.031^{+0.016}_{-0.018}. \tag{8.40}$$

The derivative improves the fit somewhat, but it is not very significantly nonzero [6], so inflation is not yet in danger. The WMAP's Hubble constant agrees perfectly with the HST value in Equation (1.20). The combined value of all present information on the Hubble constant is

$$h = 0.71^{+0.04}_{-0.03}, \tag{8.41}$$

and that on the density parameters is

$$\Omega_m h^2 = 0.135^{+0.008}_{-0.009}, \qquad \Omega_b h^2 = 0.0224 \pm 0.0009 \qquad \Omega_\nu h^2 < 0.0076. \tag{8.42}$$

It is a remarkable success of the FLRW concordance model that the baryonic density at time 380 kyr as evidenced by the CMB is in excellent agreement with the BBN evidence (5.105) from about 20 minutes after the Big Bang. As explained in Section 5.6, the BBN value depends only on the expansion rate and the nuclear reaction cross-sections, and not at all on the details of the FLRW model.

Density Parameters. From the above parameter values one derives

$$\Omega_0 = 1.02 \pm 0.02, \qquad \Omega_m = 0.27 \pm 0.04, \qquad \Omega_b = 0.044 \pm 0.004, \qquad \Omega_\nu < 0.015. \tag{8.43}$$

Thus the geometry of the Universe is consistent with being spatially flat and we can henceforth set $\Omega_0 = 1$.

Combining the limit for $\Omega_\nu h^2$ with the expression (8.16), the average mass of the three ν mass eigenstates is

$$\langle m_\nu \rangle < 0.23 \text{ eV at } 95\% \text{ CL}, \tag{8.44}$$

where 'CL' denotes the confidence level.

Inserting the values of Ω_b and h in Equation (5.102) we obtain the ratio of baryons to photons

$$\eta = (6.1 \pm 0.7) \times 10^{-10}. \tag{8.45}$$

Inserting the value of h in Equation (8.12), we obtain the present value of the radiation density parameter,

$$\Omega_r = 4.902 \times 10^{-5}. \tag{8.46}$$

Dark Energy. The equation of state of dark energy, w_φ, introduces a new degeneracy with Ω_m and h which cannot be resolved by CMB data alone. Using the full set of data, one can state a limit to w_φ (at 95% CL). The properties of dark energy are then

$$\Omega_\lambda = 0.73 \pm 0.04, \qquad w_\varphi < -0.78 \text{ at } 95\% \text{ CL}. \tag{8.47}$$

This value of Ω_λ agrees perfectly with the evidence for acceleration from supernovae in Equation (4.79). This is another remarkable success of the FLRW concordance model, since the supernova observations do not depend on Einstein's equations at all.

Similar limits on w_φ have also been obtained in independent observations. R. Jimenez *et al*. [13] combined the absolute ages of Galactic stars with the position of the first peak in the WMAP angular power spectrum, finding $w_\varphi < -0.8$ at 95% confidence. The High-z Supernova Search Team [14] have discovered and observed eight supernovae in the redshift interval $z = 0.3$–1.2, which gave them, assuming flatness, $w_\varphi < -0.73$ at 95% confidence.

Timescales. From this value for Ω_r one can determine the time of equality of radiation and matter density, t_eq. From Equations (4.13) and (4.29), the equation determining the evolution of the scale is

$$H(a)^2 = H_0^2[(1-\Omega_0)a^{-2} + \Omega(a)] = H_0^2[(1-\Omega_0)a^{-2} + \Omega_\mathrm{m}a^{-3} + \Omega_\mathrm{r}a^{-4} + \Omega_\lambda].$$

At the present time ($a = 1$) $\Omega_\mathrm{m} > \Omega_\mathrm{r}$ but, as we move back in time and a gets smaller, the term $\Omega_\mathrm{r}a^{-4}$ will come to dominate. The epoch of matter–radiation equality would have occurred when $\Omega_\mathrm{m}a^{-3} = \Omega_\mathrm{r}a^{-4}$. Using the value of $\Omega_\mathrm{m} = 0.27$ from (8.43) and Ω_r from (8.47) would give $a^{-1} = 1 + z \approx 5500$. This is not correct, however, because the a dependence of Ω_r should actually be given by

$$\Omega_\mathrm{r}(a) = \frac{g_*(a)}{2}\frac{a_\mathrm{S}T^4}{\rho_\mathrm{c}} = \frac{g_*(a)}{2}\frac{a_\mathrm{S}}{\rho_\mathrm{c}}\left(\frac{2.725}{a}\right)^4, \tag{8.48}$$

using the function g_* discussed in Chapter 5. In the region of $1 + z \gtrsim 1000$ neutrinos will be relativistic and $g_* = 3.36$ instead of 2 (the contribution to the integral (4.56) from large $z \gtrsim 10^8$, where $g_*(a) > 3.36$ is negligible). This gives

$$a_\mathrm{eq}^{-1} = 1 + z_\mathrm{eq} = \frac{0.27}{8.22 \times 10^{-5}} \approx 3300. \tag{8.49}$$

Using Equation (4.56) and $\Omega_\lambda = 0.73$ from (8.46) gives

$$t_\mathrm{eq} \approx 54\,500 \text{ yr}. \tag{8.50}$$

In Figure 5.1 we have plotted the energy density for matter and radiation as a function of the scale a of the Universe from $\log a = -6$. In Figure 5.9 we have plotted it from $\log a = -4$ until now.

Temperature scales inversely with a, thus at a_eq we have

$$T_\mathrm{eq} = T_0 a_\mathrm{eq}^{-1} \approx 9000^\circ \text{ K}. \tag{8.51}$$

Inserting the values of the energy densities (8.43) and (8.45) into Equation (4.56), one finds the age of the Universe to a remarkable precision,

$$t_0 = 13.7 \pm 0.2 \text{ Gyr}, \tag{8.52}$$

which is in excellent agreement with the independent determinations of lesser precision in Section 1.5. The WMAP team have also derived the redshift and age

of the Universe at last scattering and the thickness of the last scattering shell (noting that the WMAP team use the term *decoupling* where we have used *last scattering surface*—see our definitions in Section 5.5):

$$\left.\begin{aligned} t_{\rm LSS} &= 0.379^{+0.008}_{-0.007}\ {\rm Myr}, & \Delta t_{\rm LSS} &= 0.118^{+0.003}_{-0.002}\ {\rm Myr}, \\ 1 + z_{\rm LSS} &= 1089 \pm 1, & \Delta z_{\rm LSS} &= 195 \pm 2. \end{aligned}\right\} \tag{8.53}$$

Deceleration Parameter. Recalling the definitions

$$H_0 = \frac{\dot{R}_0}{R_0}, \qquad \Omega_{\rm m} = \frac{8\pi G\rho_{\rm m}}{3H_0^2}, \qquad \Omega_\lambda = \frac{\lambda}{3H_0^2}, \qquad q_0 = -\frac{\ddot{R}_0}{R_0 H_0^2},$$

and ignoring $\Omega_{\rm r}$, since it is so small, we can find relations between the dynamical parameters Ω_λ, $\Omega_{\rm m}$, H_0, and the deceleration parameter q_0. Substitution of these parameters into Equations (4.17) and (4.18) at present time t_0, gives

$$H_0^2 + \frac{kc^2}{R_0^2} - \frac{\lambda}{3} = \Omega_{\rm m} H_0^2, \tag{8.54}$$

$$-2q_0 H_0^2 + H_0^2 + \frac{kc^2}{R_0^2} - \lambda = -3\Omega_{\rm m} H_0^2 w, \tag{8.55}$$

where w denotes the equation of state $p_{\rm m}/\rho_{\rm m}c^2$ of matter. We can then obtain two useful relations by eliminating either k or λ. In the first case we find

$$\Omega_{\rm m}(1 + 3w) = 2q_0 + 2\Omega_\lambda, \tag{8.56}$$

and, in the second case,

$$\tfrac{3}{2}\Omega_{\rm m}(1 + w) - q_0 - 1 = \frac{kc^2}{S_0^2 H_0^2}. \tag{8.57}$$

In the present matter-dominated Universe, the pressure $p_{\rm m}$ is completely negligible and we can set $w = 0$. From Equation (8.56) and the values for $\Omega_{\rm m}$ and Ω_λ above, we find

$$q_0 = -0.60 \pm 0.02. \tag{8.58}$$

The reason for the very small error is that the errors of $\Omega_{\rm m}$ and Ω_λ are completely anti-correlated. The parameter values given in this section have been used to construct the scales in Figure 5.9. The values are collected in Table A.6 in the Appendix.

Problems

1. Derive an equation for $T_{\rm eq}$ from condition (8.3).
2. Use Wien's constant, Equation (5.106) and the CMB temperature to determine the wavelength of the CMB.

3. Use the present radiation-energy density to calculate the pressure due to radiation in the Universe. Compare this with the pressure due to a gas of galaxies calculated in Problem 3 of Chapter 3.
4. Show that an observer moving with velocity β in a direction θ relative to the CMB sees the rest frame blackbody spectrum with temperature T as a blackbody spectrum with temperature

$$T' = \frac{T}{\gamma(1 - \beta\cos\theta)}. \tag{8.59}$$

   To first order in β this gives the dipole anisotropy Equation (8.17) [1].
5. The dipole anisotropy is measured to be $1.2 \times 10^{-3} T_0$. Derive the velocity of Earth relative to the comoving coordinate system.

Chapter Bibliography

[1] Peebles, P. J. E. 1993 *Principles of physical cosmology*. Princeton University Press, Princeton, NJ.
[2] Mather, J. C., Cheng, E. S., Eplee, R. E. *et al*. 1990 *Astrophys. J. Lett.* **354**, L37.
[3] Fixsen, D. J. *et al*. 1996 *Astrophys. J.* **473**, 576.
[4] Bennett, C. L. *et al*. 1996 *Astrophys. J. Lett.* **464**, L1.
[5] Lynden-Bell, D. *et al*. 1988 *Astrophys. J.* **326**, 19.
[6] Bennett, C. L. *et al*. 2003 Preprint arXiv, astro-ph/0302207 and 2003 *Astrophys. J.* (In press.) and companion papers cited therein.
[7] Smoot, G. and Davidson, K. 1993 *Wrinkles in time*. Avon Books, New York.
[8] Kosowsky, A. 1999 *New Astronom. Rev.* **43**, 157.
[9] Kovac, J. *et al*. 2002 *Nature* **420**, 772.
[10] Colless, M. *et al*. 2001 *Mon. Not. R. Astron. Soc.* **328**, 1039.
[11] Pearson, T. J. *et al*. 2003 *Astrophys. J.* **591**, 556.
[12] Kuo, C. L. *et al*. 2002 Preprint arXiv, astro-ph/0212289.
[13] Jimenez, R., Verde, L., Treu, T. and Stern, D. 2003 Preprint arXiv, astro-ph/0302560.
[14] Tonry, J. L. *et al*. 2002 Preprint arXiv, astro-ph/0305008 and *Astrophys. J.* (2003).

9

Cosmic Structures and Dark Matter

After the decoupling of matter and radiation described in Chapter 5, we followed the fate of the free-streaming CMB in Chapter 8. Here we shall turn to the fate of matter and cold nonradiating dust. After recombination, when atoms formed, density perturbations in baryonic matter could start to grow and form structures, but growth in weakly interacting nonbaryonic (dark) matter could have started earlier at the time of radiation and matter equality. The time and size scales are important constraints to galaxy-formation models, as are the observations of curious patterns of filaments, sheets and voids on very large scales.

In Section 9.1 we describe the theory of density fluctuations in a viscous fluid, which approximately describes the hot gravitating plasma. This very much parallels the treatment of the fluctuations in radiation that cause anisotropies in the CMB.

In Section 9.2 we learn how pressure and gravitation conspire so that the hot matter can begin to cluster, ultimately to form the perhaps 10^9 galaxies, clusters and other large-scale structures.

In Section 9.3 we turn to the dynamical evidence at various scales that a large fraction of the gravitating mass in the Universe is nonluminous and composed of some unknown kind of nonbaryonic matter.

Section 9.4 lists the possible candidates of this dark matter. As we shall see, there are no good candidates, only some hypothetical particles which belong to speculative theories.

Introduction to Cosmology Third Edition by Matts Roos
© 2003 John Wiley & Sons, Ltd ISBN 0 470 84909 6 (cased) ISBN 0 470 84910 X (pbk)

In Section 9.5 we turn to observations of galaxy distributions and comparisons with simulations based on the cold dark matter (CDM) paradigm and to predictions and verifications based on the CDM paradigm.

9.1 Density Fluctuations

Until now we have described the dynamics of the Universe by assuming homogeneity and adiabaticity. The homogeneity cannot have grown out of primeval chaos, because a chaotic universe can grow homogeneous only if the initial conditions are incredibly well fine-tuned. Vice versa, a homogeneous universe will grow more chaotic, because the standard model is gravitationally unstable.

But the Universe appears homogeneous only on the largest scales (a debatable issue!), since on smaller scale we observe matter to be distributed in galaxies, groups of galaxies, supergalaxies and strings of supergalaxies with great voids in between. At the time of matter and radiation equality, some lumpiness in the energy density must have been the 'seeds' or *progenitors* of these cosmic structures, and one would expect to see traces of that lumpiness also in the CMB temperature anisotropies originating in the last scattering. The angular scale subtended by progenitors corresponding to the largest cosmic structures known, of size perhaps $200h^{-1}$ Mpc, is of the order of 3°, corresponding to CMB multipoles around $\ell = 20$.

Viscous Fluid Approximation. The common approach to the physics of matter in the Universe is by the hydrodynamics of a viscous, nonstatic fluid. With this nonrelativistic (Newtonian) treatment and linear perturbation theory we can extract much of the essential physics while avoiding the necessity of solving the full equations of general relativity. In such a fluid there naturally appear random fluctuations around the mean density $\bar{\rho}(t)$, manifested by compressions in some regions and rarefactions in other regions. An ordinary fluid is dominated by the material pressure but, in the fluid of our Universe, three effects are competing: radiation pressure, gravitational attraction and density dilution due to the Hubble flow. This makes the physics different from ordinary hydrodynamics: regions of overdensity are gravitationally amplified and may, if time permits, grow into large inhomogeneities, depleting adjacent regions of underdensity.

The nonrelativistic dynamics of a compressible fluid under gravity is described by three differential equations, the *Eulerian equations*. Let us denote the density of the fluid by ρ, the pressure p, and the velocity field $\boldsymbol{v}$, and use comoving coordinates, thus following the time evolution of a given volume of space. The first equation describes the conservation of mass: what flows out in unit time corresponds to the same decrease of matter in unit space. This is written

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = -\rho \nabla \cdot \boldsymbol{v}. \tag{9.1}$$

Next we have the equation of motion of the volume element under consideration,

$$\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = -\frac{1}{\rho}\nabla p - \nabla\phi, \tag{9.2}$$

where ϕ is the gravitational potential obeying Poisson's equation, which we met in Equation (2.85),

$$\nabla^2\phi = 4\pi G\rho. \tag{9.3}$$

Equation (9.2) shows that the velocity field changes when it encounters pressure gradients or gravity gradients.

The description in terms of the Eulerian equations is entirely classical and the gravitational potential is Newtonian. The Hubble flow is entered as a perturbation to the zeroth-order solutions with infinitesimal increments $\delta\boldsymbol{v}$, $\delta\rho$, δp and $\delta\phi$. Let us denote the local density $\rho(\boldsymbol{r}, t)$ at comoving spatial coordinate $\boldsymbol{r}$ and world time t. Then the fractional departure at $\boldsymbol{r}$ from the spatial mean density $\bar{\rho}(t)$ is the dimensionless *mass density contrast*

$$\delta_{\mathrm{m}}(\boldsymbol{r}, t) = \frac{\rho_{\mathrm{m}}(\boldsymbol{r}, t) - \bar{\rho}_m(t)}{\bar{\rho}_m(t)}. \tag{9.4}$$

The solution to Equations (9.1)–(9.3) can then be sought in the form of waves,

$$\delta_{\mathrm{m}}(\boldsymbol{r}, t) \propto \mathrm{e}^{\mathrm{i}(\boldsymbol{k}\cdot\boldsymbol{r} - \omega t)}, \tag{9.5}$$

where $\boldsymbol{k}$ is the wave vector in comoving coordinates. An arbitrary pattern of fluctuations can be described mathematically by an infinite sum of independent waves, each with its characteristic wavelength λ or comoving wavenumber k and its amplitude δ_k. The sum can be formally expressed as a Fourier expansion for the density contrast

$$\delta_{\mathrm{m}}(\boldsymbol{r}, t) \propto \sum \delta_k(t)\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}. \tag{9.6}$$

A density fluctuation can also be expressed in terms of the mass M moved within one wavelength, or rather within a sphere of radius λ, thus $M \propto \lambda^3$. It follows that the wavenumber or spatial frequency k depends on mass as

$$k = \frac{2\pi}{\lambda} \propto M^{-1/3}. \tag{9.7}$$

Power Spectrum. The density fluctuations can be specified by the amplitudes δ_k of the dimensionless *mass autocorrelation function*

$$\xi(r) = \langle\delta(\boldsymbol{r}_1)\delta(\boldsymbol{r} + \boldsymbol{r}_1)\rangle \propto \sum\langle|\delta_k(t)|^2\rangle\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}, \tag{9.8}$$

which measures the correlation between the density contrasts at two points $\boldsymbol{r}$ and $\boldsymbol{r}_1$. The powers $|\delta_k|^2$ define the power spectrum of the root-mean-squared (RMS) mass fluctuations

$$P(k) = \langle|\delta_k(t)|^2\rangle. \tag{9.9}$$

Thus the autocorrelation function $\xi(r)$ is the Fourier transform of the power spectrum. We have already met a similar situation in the context of CMB anisotropies, where the waves represented temperature fluctuations on the surface of the surrounding sky. There Equation (8.18) defined the autocorrelation function $C(\theta)$

and the powers a_ℓ^2 were coefficients in the Legendre polynomial expansion Equation (8.19).

Taking the power spectrum to be of the phenomenological form (7.57),

$$P(k) \propto k^n,$$

and combining with Equation (9.7), one sees that each mode δ_k is proportional to some power of the characteristic mass enclosed, M^α.

Inflationary models predict that the mass density contrast obeys

$$\delta_{\mathrm{m}}^2 \propto k^3 \langle |\delta_k(t)|^2 \rangle \tag{9.10}$$

and that the primordial fluctuations have approximately a Harrison–Zel'dovich spectrum with $n_{\mathrm{s}} = 1$. Support for these predictions come from the CMB temperature and polarization asymmetry spectra which give the value quoted in Equation (8.40), $n = 0.93 \pm 0.03$ [1].

Independent, although less accurate information about the spectral index can be derived from constraints set by CMB isotropy, galaxies and black holes using $\delta_k \propto M^\alpha$. The CMB scale (within the size of the present horizon of $M \approx 10^{22} M_\odot$) is isotropic to less than about 10^{-4}. Galaxy formation (scale roughly $10^{12} M_\odot$) requires perturbations of order $10^{-4\pm1}$. Taking the ratio of perturbations versus mass implies a constraint on α and implies that n is close to 1.0 at long wavelengths.

Turning to short wavelengths (scale about 10^{12} kg or $10^{-18} M_\odot$), black holes provide another constraint. Primordial perturbations on this scale must have been roughly smaller than 1.0. Larger perturbations would have led to production of many black holes, since large-amplitude perturbations inevitably cause overdense regions to collapse before pressure forces can respond. From Equation (3.30) one sees that black holes less massive than 10^{12} kg will have already evaporated within 10 Gyr, but those more massive will remain and those of mass 10^{12} kg will be evaporating and emitting γ-rays today. Large amplitude perturbations at and above 10^{12} kg would imply more black holes than is consistent with the mass density of the Universe and the γ-ray background. Combining the black hole limit on perturbations (up to around 1 for $M \approx 10^{12}$ kg) with those from the CMB and galaxy formation also implies the spectrum must be close to the Harrison–Zel'dovich form.

The power spectra of theoretical models for density fluctuations can be compared with the real distribution of galaxies and galaxy clusters. Suppose that the galaxy number density in a volume element $\mathrm{d}V$ is n_{G}, then one can define the probability of finding a galaxy in a random element as

$$\mathrm{d}P = n_{\mathrm{G}}\, \mathrm{d}V. \tag{9.11}$$

If the galaxies are distributed independently, for instance with a spatially homogeneous Poisson distribution, the joint probability of having one galaxy in each of two random volume elements $\mathrm{d}V_1$, $\mathrm{d}V_2$ is

$$\mathrm{d}P_{12} = n_{\mathrm{G}}^2\, \mathrm{d}V_1\, \mathrm{d}V_2. \tag{9.12}$$

There is then no correlation between the probabilities in the two elements. However, if the galaxies are clustered on a characteristic length r_c, the probabilities in different elements are no longer independent but correlated. The joint probability of having two galaxies with a relative separation r can then be written

$$dP_{12} = n_G^2[1 + \xi(r/r_c)]\, dV_1\, dV_2, \tag{9.13}$$

where $\xi(r/r_c)$ is the *two-point correlation function* for the galaxy distribution. This can be compared with the autocorrelation function (9.8) of the theoretical model. If we choose our own Galaxy at the origin of a spherically symmetric galaxy distribution, we can simplify Equation (9.13) by setting $n_G\, dV_1 = 1$. The right-hand side then gives the average number of galaxies in a volume element dV_2 at distance r.

Analyses of galaxy clustering show [2] that, for distances

$$10 \text{ kpc} \lesssim hr \lesssim 10 \text{ Mpc}, \tag{9.14}$$

a good empirical form for the two-point correlation function is

$$\xi(r/r_c) = (r/r_c)^{-\gamma}, \tag{9.15}$$

with the parameter values $r_c \approx 5.0h^{-1}$ Mpc, $\gamma \approx 1.7$.

Irregularities in the metric can be expressed by the curvature radius r_U defined in Equation (4.54). If r_U is less than the linear dimensions d of the fluctuating region, it will collapse as a black hole. Establishing the relation between the curvature of the metric and the size of the associated mass fluctuation requires the full machinery of general relativity, which is beyond our ambitions.

Linear Approximation. Much of the interesting physics of density fluctuations can be captured by a Newtonian linear perturbation analysis of a viscous fluid. Small perturbations grow slowly over time and follow the background expansion until they become heavy enough to separate from it and to collapse into gravitationally bound systems. As long as these perturbations are small they can be decomposed into Fourier components that develop independently and can be treated separately. For fluctuations in the linear regime, $|\delta_k| < 1$, where

$$\rho_m = \bar{\rho}_m + \Delta\rho_m, \qquad p = \bar{p} + \Delta p, \qquad v^i = \bar{v}^i + \Delta v^i, \qquad \phi = \bar{\phi} + \Delta\phi, \tag{9.16}$$

the size of the fluctuations and the wavelengths grows linearly with the scale a, whereas in the nonlinear regime, $|\delta_k| > 1$, the density fluctuations grow faster, with the power a^3, at least (but not exponentially). The density contrast can also be expressed in terms of the linear size d of the region of overdensity normalized to the curvature radius,

$$\delta \approx \left(\frac{d}{r_U}\right)^2. \tag{9.17}$$

In the linear regime r_U is large, so the Universe is flat. At the epoch when d is of the order of the Hubble radius, the density contrast is

$$\delta_H \approx \left(\frac{r_H}{r_U}\right)^2, \tag{9.18}$$

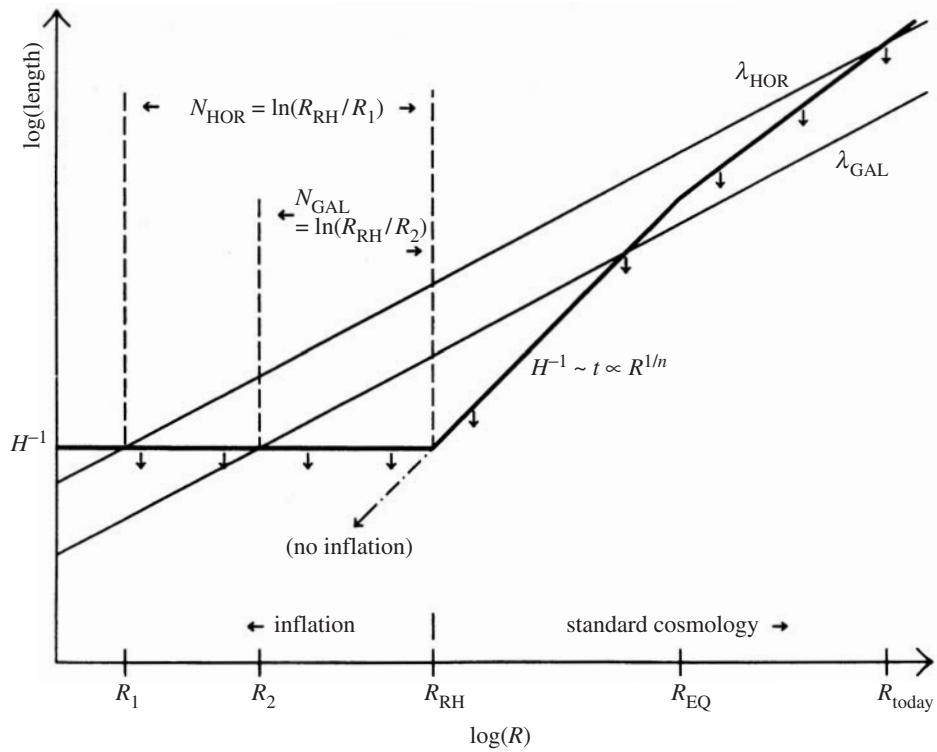


Figure 9.1 The evolution of the physical size of the comoving scale or wavelength λ, and of the Hubble radius H^{-1} as functions of the scale of the Universe R. In the standard noninflationary cosmology, a given scale crosses the horizon but once, while in the inflationary cosmology all scales begin sub-horizon sized, cross outside the Hubble radius ('good bye') during inflation, and re-enter ('hello again') during the post-inflationary epoch. Note that the largest scales cross outside the Hubble radius first and re-enter last. The growth in the scale factor, $N = \ln(R_{RH}/R)$, between the time a scale crosses outside the Hubble radius during inflation and the end of inflation is also indicated. For a galaxy, $N_{GAL} = \ln(R_{RH}/R_2) \approx 45$, and for the present horizon scale, $N_{HOR} = \ln(R_{RH}/R_1) \approx 53$. Causal microphysics operates only on scales less than H^{-1}, indicated by arrows. During inflation H^{-1} is a constant, and in the post-inflation era it is proportional to $R^{1/n}$, where $n = 2$ during radiation domination, and $n = \frac{3}{2}$ during matter domination. Courtesy of E. W. Kolb and M. S. Turner.

free streaming can leave the region and produce the CMB anisotropies. Structures formed when $d \ll r_H$, thus when $\delta \ll 1$. Although δ may be very small, the fluctuations may have grown by a very large factor because they started early on (see Problem 3 in Chapter 7).

When the wavelength is below the horizon, causal physical processes can act and the (Newtonian) viscous fluid approximation is appropriate. When the wavelength is of the order of or larger than the horizon, however, the Newtonian analysis is not sufficient. We must then use general relativity and the choice of gauge is important.

Gauge Problem. The mass density contrast introduced in Equation (9.4) and the power spectrum of mass fluctuations in Equation (9.9) represented perturbations to an idealized world, homogeneous, isotropic, adiabatic, and described by the FLRW model. For sub-horizon modes this is adequate. For super-horizon modes one must apply a full general-relativistic analysis. Let us call the space-time of the world just described $\mathcal{G}$. In the real world, matter is distributed as a smooth background with mass perturbations imposed on it. The space-time of that world is not identical to $\mathcal{G}$, so let us call it $\mathcal{G}'$.

To go from $\mathcal{G}'$, where measurements are made, to $\mathcal{G}$, where the theories are defined, requires a *gauge transformation*. This is something more than a mere coordinate transformation—it also changes the event in $\mathcal{G}$ that is associated to an event in $\mathcal{G}'$. A perturbation in a particular observable is, by definition, the difference between its value at some space-time event in $\mathcal{G}$ and its value at the corresponding event in the background (also in $\mathcal{G}$). An example is the mass auto-correlation function $\xi(r)$ in Equation (9.8).

But this difference need not be the same in $\mathcal{G}'$. For instance, even if an observable behaves as a scalar under coordinate transformations in $\mathcal{G}$, its perturbation will not be invariant under gauge transformations if it is time dependent in the background. Non-Newtonian density perturbations in $\mathcal{G}$ on super-horizon scales may have an entirely different time dependence in $\mathcal{G}'$, and the choice of gauge transformation $\mathcal{G} \to \mathcal{G}'$ is quite arbitrary.

But arbitrariness does not imply that one gauge is correct and all others wrong. Rather, it imposes on physicists the requirement to agree on a convention, otherwise there will be problems in the interpretation of results. The formalism we chose in Chapter 2, which led to Einstein's equation (2.96) and Friedmann's equations (4.4) and (4.5), implicitly used a conventional gauge. Alternatively one could have used gauge-invariant variables, but at the cost of a very heavy mathematical apparatus. Another example which we met briefly in Section 6.2 concerned the electroweak theory, in which particle states are represented by gauge fields that are locally gauged.

9.2 Structure Formation

As we have seen in the FLRW model, the force of gravity makes a homogeneous matter distribution unstable: it either expands or contracts. This is true for matter on all scales, whether we are considering the whole Universe or a tiny localized region. But the FLRW expansion of the Universe as a whole is not exponential and therefore it is too slow to produce our Universe in the available time. This requires a different mechanism to give the necessary exponential growth: cosmic inflation. Only after the graceful exit from inflation does the Universe enter the regime of Friedmann expansion, during which the Hubble radius gradually overtakes the inflated regions. Thus, inflationary fluctuations will cross the post-inflationary Hubble radius and come back into vision with a wavelength corresponding to the size of the Hubble radius at that moment. This is illustrated in Figure 9.1.

Jeans Mass. Primordial density fluctuations expand linearly at a rate slower than the Universe is expanding in the mean, until eventually they reach a maximum size and collapse nonlinearly. If the density fluctuates locally, also the cosmic scale factor will be a fluctuating function $a(\boldsymbol{r}, t)$ of position and time. In overdense regions where the gravitational forces dominate over pressure forces, causing matter to contract locally and to attract surrounding matter which can be seen as inflow. In other regions where the pressure forces dominate, the fluctuations move as sound waves in the fluid, transporting energy from one region of space to another.

The dividing line between these two possibilities can be found by a classical argument. Let the time of free fall to the centre of an inhomogeneity in a gravitational field of strength G be

$$t_{\mathrm{G}} = 1/\sqrt{G\bar{\rho}}. \tag{9.19}$$

Sound waves in a medium of density ρ and pressure p propagate with velocity

$$c_{\mathrm{s}} = \sqrt{\frac{\partial p}{\partial \rho}},$$

so they move one wavelength λ in the time

$$t_{\mathrm{s}} = \lambda/c_{\mathrm{s}}. \tag{9.20}$$

Note that only baryonic matter experiences pressure forces. In the next sections we shall meet noninteracting forms of matter which feel only gravitational forces.

When t_{G} is shorter than t_{s}, the fluctuations are unstable and their amplitude will grow by attracting surrounding matter, becoming increasingly unstable until the matter eventually collapses into a gravitationally bound object. The opposite case is stable: the fluctuations will move with constant amplitude as sound waves. Setting $t_{\mathrm{G}} = t_{\mathrm{s}}$, we find the limiting *Jeans wavelength* $\lambda = \lambda_{\mathrm{J}}$ at the *Jeans instability*, discovered by *Sir James Jeans* (1877–1946) in 1902,

$$\lambda_{\mathrm{J}} = \sqrt{\frac{\pi}{G\bar{\rho}}}\, c_{\mathrm{s}}. \tag{9.21}$$

Actually, the factor $\sqrt{\pi}$ was not present in the above Newtonian derivation; it comes from an exact treatment of Equations (9.1)–(9.3) (see for instance reference [3]). The mass contained in a sphere of radius λ_{J} is called the *Jeans mass*,

$$M_{\mathrm{J}} = \tfrac{4}{3}\pi\lambda_{\mathrm{J}}^3\rho. \tag{9.22}$$

In order for tiny density fluctuations to be able to grow to galactic size, there must be enough time, or the expansion must be exponential in a brief time. The gravitational collapse of a cloud exceeding the Jeans mass develops exponentially, so the cloud halves its radius in equal successive time intervals. But galaxies and large-scale structures do not condense out of the primordial medium by exponential collapse. The structures grow only linearly with the scale a or as some low power of a.

For sub-horizon modes, the distinction between the radiation-dominated and matter-dominated eras is critical. During the radiation era, growth of perturbations is suppressed. During the matter era, perturbations can grow. But during the matter era the Jeans wavelength provides an important boundary. Large wavelength fluctuations will grow with the expansion as long as they are in the linear regime. In an accelerated expansion driven by dark energy, the condition for gravitational collapse becomes extremely complicated. This happens rather late, only when matter domination ends and dark energy becomes dynamically important ($z \sim 1$).

For wavelengths less than the Jeans wavelength the pressure in the baryonic matter can oppose the gravitational collapse and perturbations will oscillate in the linear regime as sound waves, never reaching gravitational collapse. An alternative way of stating this is to note that the radiation pressure and the tight coupling of photons, protons and electrons causes the fluid to be viscous. On small scales, photon diffusion and thermal conductivity inhibit the growth of perturbations as soon as they arise, and on large scales there is no coherent energy transport.

Mass fluctuations at still shorter wavelength, with $\lambda \approx r_{\rm U} \ll r_{\rm H}$, can break away from the general expansion and collapse to bound systems of the size of galaxies or clusters of galaxies. Fluctuations which enter in the nonlinear regime, where the ratio in Equation (9.17) is large, collapse rapidly into black holes before pressure forces have time to respond.

For baryonic matter before the recombination era, the baryonic Jeans mass is some 30 times larger than the mass $M_{\rm H}$ of baryons within the Hubble radius $r_{\rm H}$, so if there exist nonlinear modes they are outside it (the Jeans wavelength is greater than the horizon). A mass scale M is said to enter the Hubble radius when $M = M_{\rm H}$. Well inside the Hubble radius, the fluctuations may start to grow as soon as the Universe becomes matter dominated, which occurs at time $t_{\rm eq} = 54\,500$ yr (from Equation (8.50)).

Upon recombination, the baryonic Jeans mass falls dramatically. If the fluid is composed of some nonbaryonic particle species (cold dark matter), the Jeans wavelength is small after radiation–matter equality, allowing sub-horizon perturbations to grow from this time. After matter–radiation equality, nonbaryonic matter can form potential wells into which baryons can fall after recombination.

Matter can have two other effects on perturbations. Adiabatic fluctuations lead to gravitational collapse if the mass scale is so large that the radiation does not have time to diffuse out of one Jeans wavelength within the time $t_{\rm eq}$. As the Universe approaches decoupling, the photon mean free path increases and radiation can diffuse from overdense regions to underdense ones, thereby smoothing out any inhomogeneities in the plasma. For wavelengths below the Jeans wavelength, *collisional dissipation* or *Silk damping* (after *J. Silk*) erases perturbations in the matter (baryon) radiation field through photon diffusion. This becomes most important around the time of recombination. Random waves moving through the medium with the speed of sound $c_{\rm s}$ erase all perturbations with wavelengths less than $c_{\rm s} t_{\rm eq}$. This mechanism sets a lower limit to the size of the structures that can form by the time of recombination: they are not smaller than rich clusters

or superclusters. But, in the presence of nonbaryonic matter, Silk damping is of limited importance because nonbaryonic matter does not couple with the radiation field.

The second effect is free streaming of weakly interacting relativistic particles such as neutrinos. This erases perturbations up to the scale of the horizon, but this also ceases to be important at the time of matter–radiation equality.

The situation changes dramatically at recombination, when all the free electrons suddenly disappear, captured into atomic Bohr orbits, and the radiation pressure almost vanishes. This occurs at time 400 000 yr after Big Bang (see Figure 5.9). Now the density perturbations which have entered the Hubble radius can grow with full vigour.

Sunyaev–Zel'dovich Effect (SZE). At some stage the hydrogen gas in gravitationally contracting clouds heats up enough to become ionized and to re-ionize the CMB: the Sunyaev–Zel'dovich effect. We refer to the WMAP result in Equation (8.39) that such re-ionization clouds occur at an average redshift $z_r \approx 20$.

The free electrons and photons in the ionized clouds build up a radiation pressure, halting further collapse. The state of such clouds today depends on how much mass and time there was available for their formation. Small clouds may shrink rapidly, radiating their gravitational binding energy and fragmenting. Large clouds shrink slowly and cool by the mechanism of electron Thomson scattering. As the recombination temperature is approached the photon mean free paths become larger, so that radiation can diffuse out of overdense regions. This damps the growth of inhomogeneities.

The distortion of the CMB spectrum due to the SZE can be used to detect intergalactic clouds and to provide another estimate of H_0 by combining radio and X-ray observations to obtain the distance to the cloud. The importance of the SZE surveys is that they are able to detect all clusters above a certain mass limit independent of the redshifts of the clusters. The ratio of the magnitude of the SZE to the CMB does not change with redshift. The effects of re-ionization on the CMB temperature–polarization power were discussed in Section 8.4.

Structure Sizes and Formation Times. Only clouds exceeding the Jeans mass stabilize and finally attain *virial equilibrium*. It is intriguing (but perhaps an accident) that the Jeans mass just after recombination is about $10^5 M_\odot$, the size of globular clusters! Galaxies have masses of the order of $10^{12} M_\odot$ corresponding to fluctuations of order $\delta \simeq 10^{-4}$ as they cross the horizon. We have already made use of this fact to fix the mass m_φ of the scalar field in Equation (7.49).

The timetable for galaxy and cluster formation is restricted by two important constraints. At the very earliest, the Universe has to be large enough to have space for the first formed structures. If these were galaxies of the present size, their number density can be used to estimate how early they could have been formed. We leave this for a problem.

The present density of the structures also sets a limit on the formation time. The density contrast at formation must have exceeded the mean density at that time, and since then the contrast has increased with a^3. Thus, rich clusters, for instance, cannot have been formed much earlier than at

$$1 + z \approx 2.5\Omega^{-1/3}. \tag{9.23}$$

It seems that all the present structure was already in place at $z = 5$. This does not exclude the fact that the largest clusters are still collapsing today. In a critical universe structure formation occurs continuously, rich galaxy clusters form only at a redshift of 0.2–0.3, and continue to accrete material even at the present epoch. In that case many clusters are expected to show evidence for recent merger events and to have irregular morphologies.

As a result of mass overdensities, the galaxies influenced by the ensuing fluctuations in the gravitational field will acquire peculiar velocities. One can derive a relation between the mass autocorrelation function and the RMS peculiar velocity (see reference [3]). If one takes the density contrast to be $\delta_{\rm m} = 0.3$ for RMS fluctuations of galaxy number density within a spherical volume radius $30h^{-1}$ Mpc, and if one further assumes that all mass fluctuations are uncorrelated at larger separations, then the acceleration caused by the gravity force of the mass fluctuations would predict deviations from a homogeneous peculiar velocity field in rough agreement with observations in our neighbourhood. Much larger density contrast would be in marked disagreement with the standard model and with the velocity field observations.

9.3 The Evidence for Dark Matter

Below we shall study the matter content in dynamical systems on a variety of scales: galaxies, small galaxy groups, the local supercluster, rich galaxy clusters, and the full Universe inside our horizon.

Inventory. In Equation (8.43) we quoted a value for the mass density $\Omega_{\rm m}$ from a combination of CMB, large-scale-structure and supernova data. At that point we did not specify the types of gravitating mass that $\Omega_{\rm m}$ represented—baryons certainly, and neutrinos too. Let us at this point make a full inventory of the known contents of energy densities in the Universe on a large scale. This implies rewriting Equation (4.20) in detail:

$$\Omega_0 = \Omega_{\rm b} + \Omega_\nu + \Omega_{\rm r} + \Omega_\lambda. \tag{9.24}$$

We know that

(i) $\Omega_0 = 1.02 \pm 0.02$ from Equation (8.43); this permits us to consider the Universe as being spatially flat ($\Omega_0 = 1$). Moreover,

(ii) $\Omega_{\rm b} = 0.044 \pm 0.004$ from Equation (8.43),

(iii) $\Omega_\nu < 0.015$ from Equations (8.16) and (8.43),

(iv) $\Omega_\mathrm{r} = 4.902 \times 10^{-5}$ from Equation (8.46),

(v) $\Omega_\lambda = 0.73 \pm 0.04$ from Equation (8.47).

Obviously, Ω_b does not, by far, make up all the matter content in $\Omega_\mathrm{m} = 0.27$, and Ω_ν and Ω_r can be neglected here. Therefore, a large fraction is missing from the right-hand side of (9.24) to make the equation balance. All forms of radiating baryonic mass are already accounted for in Ω_b: starlight amounts to $\Omega_* = 0.001$–0.002, gas and star remnants in galaxies amount to $\Omega_\mathrm{lum} < 0.01$. The intergalactic gas contains hydrogen clouds and filaments seen by its Lyα absorption, warm gas in groups of galaxies radiates soft X-rays, hot gas in clusters is seen in keV X-rays and in the SZE. The missing fraction is not radiating and is therefore called *dark matter* (DM)

$$\Omega_\mathrm{dm} = \Omega_\mathrm{m} - \Omega_\mathrm{b} = 0.23 \pm 0.05. \tag{9.25}$$

The remarkable fact is that the missing energy density is much larger than the known fraction.

Galaxy Formation. Galaxies form by gas cooling and condensing into DM haloes, where they turn into stars. The star-formation rate is $10 M_\odot\ \mathrm{yr}^{-1}$ in galaxies at $2.8 < z < 3.5$ for which the Lyα break at 91.2 nm shifts significantly (at $z = 3$ it has shifted to 364.8 nm). Galaxy mergers and feedback processes also play major roles.

If the galaxies have arisen from primordial density fluctuations in a purely baryonic medium, the amplitude of the fluctuations must have been very large, since the amount of baryonic matter is so small. But the amplitude of fluctuations in the CMB must then also be very large, because of adiabaticity. This leads to intolerably large CMB anisotropies today. Thus galaxy formation in purely baryonic matter is ruled out by this argument alone.

Spiral Galaxies. The spiral galaxies are stable gravitationally bound systems in which matter is composed of stars and interstellar gas. Most of the observable matter is in a relatively thin disc, where stars and gas travel around the galactic centre on nearly circular orbits. By observing the Doppler shift of the integrated starlight and the radiation at $\lambda = 0.21$ m from the interstellar hydrogen gas, one finds that galaxies rotate. If the circular velocity at radius r is v in a galaxy of mass M, the condition for stability is that the centrifugal acceleration should equal the gravitational pull:

$$\frac{v^2}{r} = \frac{GM}{r^2}. \tag{9.26}$$

In other words, the radial dependence of the velocity of matter rotating in a disc is expected to follow Kepler's law

$$v = \sqrt{\frac{GM}{r}}. \tag{9.27}$$

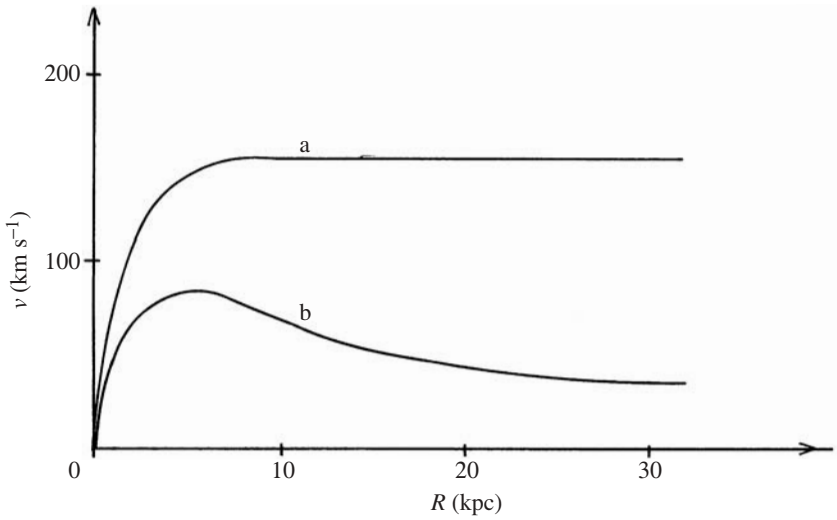


Figure 9.2 Typical galaxy rotation curves: (a) derived from the observed Doppler shift of the 0.21 m line of atomic hydrogen relative to the mean; (b) prediction from the radial light distribution.

Visible starlight traces velocity out to radial distances typically of the order of 10 kpc, and interstellar gas out to 20–50 kpc. The surprising result from measurements of galaxy-rotation curves is that the velocity does not follow the $1/\sqrt{r}$ law (9.27), but stays constant after a maximum at about 5 kpc (see Figure 9.2). Assuming that the disc-surface brightness is proportional to the surface density of luminous matter, one derives a circular speed which is typically more than a factor of three lower than the speed of the outermost measured points (see, for example, reference [4]). This implies that the calculated gravitational field is too small by a factor of 10 to account for the observed rotation. This effect is even more pronounced in *dwarf spheroidal galaxies*, which are only a few parsecs across and are the smallest systems in which dynamical DM has been detected. These galaxies have the highest known DM densities, approximately $1 M_\odot\ \mathrm{pc}^{-3}$, and the motion of their stars is completely dominated by DM at all radii.

There are only a few possible solutions to this problem. One is that the theory of gravitation is wrong. It is possible to modify ad hoc Kepler's inverse square law or Newton's assumption that G is a constant, but the corresponding modifications cannot be carried out in the relativistic theory, and a general correlation between mass and light remains. The modifications would have to be strong at large scales, and this would greatly enhance cosmic shear, which is inconsistent with measurements.

Another possibility is that spiral galaxies have magnetic fields extending out to regions of tens of kiloparsecs where the interstellar gas density is low and the gas dynamics may easily be modified by such fields [5]. But this argument works only on the gas halo, and does not affect the velocity distribution of stars. Also, the existence of magnetic fields of sufficient strength remains to be demonstrated; in our Galaxy it is only a few microgauss, which is insufficient.

The solution attracting most attention is that there exist vast amounts of non-luminous DM beyond that accounted for by stars and hydrogen clouds. One natural place to look for DM is in the neighbourhood of the Solar System. In 1922, *Jacobus C. Kapteyn* deduced that the total density in the local neighbourhood is about twice as large as the luminous density in visible stars. Although the result is somewhat dependent on how large one chooses this 'neighbourhood' to be, modern dynamical estimates are similar.

The luminous parts of galaxies, as evidenced by radiation of baryonic matter in the visible, infrared and X-ray spectra, account only for $\Omega_{\text{lum}} < 0.01$. The internal dynamics implies that the galaxies are embedded in extensive haloes of DM, of the order of

$$\Omega_{\text{halo}} > 0.03\text{–}0.10. \tag{9.28}$$

In fact, to explain the observations, the radial mass distribution $M(r)$ must be proportional to r,

$$v = \sqrt{\frac{GM(r)}{r}} \propto \sqrt{\frac{Gr}{r}} = \text{const.} \tag{9.29}$$

The radial density profile is then

$$\rho(r) \propto r^{-2}. \tag{9.30}$$

This is precisely the distribution one would obtain if the galaxies were surrounded by a halo formed by an isothermal gas sphere, where the gas pressure and gravity were in virial equilibrium.

Actually, the observed rotation curves show no obvious density plateau or core near the centre. The profile of DM in haloes is shallower than isothermal near the centre and steeper in the outer parts. The inner profiles ($r < r_{\text{s}}$) of DM haloes are remarkably similar and well approximated by a power-law central cusp of the form

$$\frac{\rho(r)}{\rho_{\text{c}}(z)} = \frac{\delta_{\text{c}}}{(r/r_{\text{s}})(1 + r/r_{\text{s}})^2}, \tag{9.31}$$

where δ_{c} and r_{s} are two free parameters to be fitted and $\rho_{\text{c}}(z)$ is the critical density at the redshift of the galaxy [6]. The outer shape varies from halo to halo, mostly because of the presence of minihalo remnants and other substructures.

Thus the unavoidable conclusion is that galactic haloes contain DM: in fact an order of magnitude more DM than baryonic matter. For instance, there is five times more DM mass than visible baryonic mass in the M33 galaxy, and the ratio is 50:1 in its halo [7].

Small Galaxy Groups. Let us now turn to gravitational systems formed by a small number of galaxies. There are examples of such groups in which the galaxies are enveloped in a large cloud of hot gas, visible by its X-ray emission. The amount of gas can be deduced from the intensity of this radiation. Adding this to the luminous matter, the total amount of baryonic matter can be estimated. The temperature of the gas depends on the strength of the gravitational field, from which the total amount of gravitating matter in the system can be deduced.

In the galaxy group HCG62 in the Coma cluster, the ROSAT satellite has found a temperature of about 10^7 K [8], which is much higher than that which the gravitational field of visible baryonic matter (galaxies and gas) would produce. One then deduces a baryonic mass fraction of $\Omega_b \gtrsim 0.13$. But this cannot be typical of the Universe as a whole, since it conflicts with the value in Equation (8.43).

The Local Supercluster (LSC). The autocorrelation function $\xi(r)$ in Equation (9.8) was defined for distances r in real space. In practice, distances to galaxies are measured in redshifts, and then two important distortions enter. To describe the separation of galaxy pairs on the surface of the sky, let us introduce the coordinate σ, transversal to the line of sight, and π radial. In redshift space the correlation function is then described by $\xi(\sigma, \pi)$ or its spherical average $\xi(s)$, where $s = \sqrt{\pi^2 + \sigma^2}$.

The transversal distance σ is always accurate, but the radial redshift distance π is affected by velocities other than the isotropic Hubble flow. For relatively nearby galaxies, $r \leqslant 2$ Mpc, the random peculiar velocities make an unknown contribution to π so that $\xi(s)$ is radially distorted. The undistorted correlation function $\xi(r)$ is seen isotropic in (σ, π)-space in the top left panel of Figure 9.3. The lower left panel of Figure 9.3 shows the distortion to $\xi(s)$ as an elongation in the π direction.

Over large distances where the peculiar velocities are unimportant relative to the Hubble flow (tens of Mpc), the galaxies in the LSC feel its attraction, as is manifested by their infall toward its centre with velocities in the range 150–450 km s^{-1}. From this one can derive the local gravitational field and the mass excess δM concentrated in the LSC. The infall velocities cause another distortion to $\xi(s)$: a flattening as is shown in the top right panel of Figure 9.3. When both distortions are included, the correlation function in (σ, π)-space looks like the bottom right panel of Figure 9.3. The narrow peaks in the π direction have been seen for a long time, and are called *Fingers of God*.

If galaxy formation is a local process, then on large scales galaxies must trace mass (on small scales galaxies are less clustered than mass), so that $\xi_{gal}(r)$ and $\xi_{mass}(r)$ are proportional:

$$\xi_{gal}(r) = b^2 \xi_{mass}(r).$$

Here b is the linear *bias*: bias is when galaxies are more clustered than mass, and *anti-bias* is the opposite case; $b = 1$ corresponds to the unbiased case. The presence of bias is an inevitable consequence of the nonlinear nature of galaxy formation. The distortions in $\xi(s)$ clearly depend on the mass density Ω_m within the observed volume. Introducing the phenomenological flattening parameter

$$\beta = \Omega_m^{0.6}/b, \tag{9.32}$$

one can write a linear approximation to the distortion as

$$\frac{\xi(s)}{\xi(r)} = 1 + \frac{2\beta}{3} + \frac{\beta^2}{5}. \tag{9.33}$$

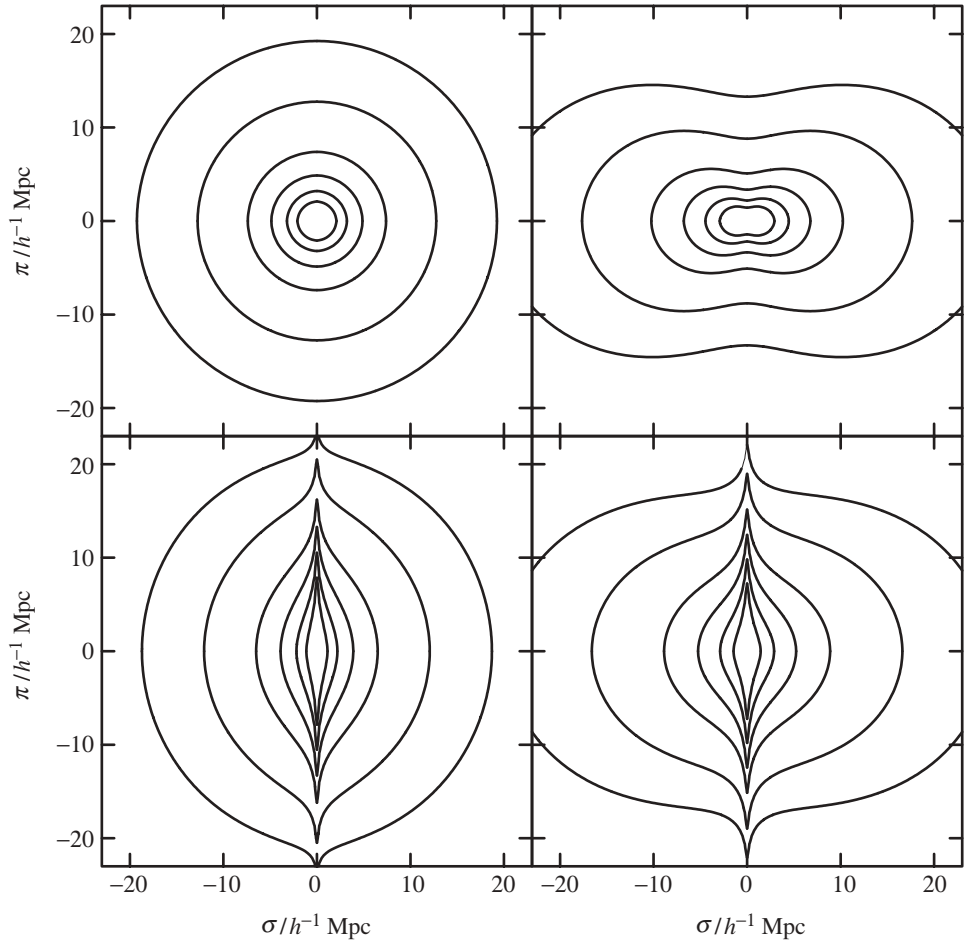


Figure 9.3 Plot of theoretically calculated correlation functions $\xi(\sigma, \pi)$ as described in reference [2]. The lines represent contours of constant $\xi(\sigma, \pi) = 4.0, 2.0, 0.5, 0.2, 0.1$. The models are: top left, undistorted; bottom left, no infall velocities but $\beta = 0.4$; top right, infall velocity dispersion $a = 500$ km s^{-1}, $\beta = 0$; bottom right, $a = 500$ km s^{-1}, $\beta = 0.4$. All models use Equation (9.15) with $r_c = 5.0h^{-1}$ Mpc and $\gamma = 1.7$. Reproduced from reference [2] by permission of the 2dFGRS Team.

Estimates of β and b which we shall discuss later lead to a value of Ω_m of the order of 0.3 [2]. Thus a large amount of matter in the LSC is dark.

Rich Clusters. The classical method for obtaining the mean mass density, ρ_m, is to measure the mean ratio of mass to luminosity, $Y = M/L$ for a representative sample of galaxies. The light radiated by galaxies is a universal constant L_U and the mass density is then obtained as $L_U Y$ (Problem 6). In solar units, $M_\odot/L_\odot$, the value for the Sun is $Y = 1$ and values in the solar neighbourhood are about 2.5–7. Similar values apply to small galaxy groups. However, rich galaxy clusters exhibit

much larger Υ values, from 300 for the Coma cluster to 650. These large values show that rich clusters have their own halo of DM which is much larger than the sum of haloes of the individual galaxies.

Zwicky noted in 1933 that the galaxies in the Coma cluster and other rich clusters move so fast that the clusters require about 10–100 times more mass to keep the galaxies bound than could be accounted for by the luminous mass in the galaxies themselves. This was the earliest indication of DM in objects at cosmological distances.

The virial theorem for a statistically steady, spherical, self-gravitating cluster of objects, stars or galaxies states that the total kinetic energy of N objects with average random peculiar velocities v equals $-\frac{1}{2}$ times the total gravitational potential energy. If r is the average separation between any two objects of average mass m, the potential energy of each of the possible $N(N-1)/2$ pairings is $-Gm^2/r$. The virial theorem then states that

$$N\frac{mv^2}{2} = \frac{1}{2}\frac{N(N-1)}{2}\frac{Gm^2}{r}. \tag{9.34}$$

For a large cluster of galaxies of total mass M and radius r, this reduces to

$$M = \frac{2rv^2}{G}. \tag{9.35}$$

Thus, one can apply the virial theorem to estimate the total dynamic mass of a rich galaxy cluster from measurements of the velocities of the member galaxies and the cluster radius from the volume they occupy. When such analyses have been carried out, taking into account that rich clusters have about as much mass in hot gas as in stars, one finds that gravitating matter accounts for

$$\Omega_{\rm grav} = 0.2\text{–}0.3, \tag{9.36}$$

or much more than the fraction of baryonic matter.

This is well demonstrated in a study of groups and clusters of galaxies with the Position Sensitive Proportional Counter (PSPC) instrument on board the X-ray satellite ROSAT [9, 10]. An example is given in Figure 9.4, which shows the radial distribution of various gravitating components in the hot cluster A85. The intracluster gas visible by its X-ray emission is the most extended mass component (GAS), and galaxies constitute the most centrally concentrated component (GAL). Thus the total baryonic matter at large radii is well approximated by GAS, and it is reasonable to assume that the constant level it approaches corresponds to the primordial composition out of which galaxies formed. One clearly sees the need for a dominating DM component (DARK) which is clustered intermediate between GAL and GAS.

One can then establish the relation

$$f_{\rm GAS} = \frac{\Omega_{\rm b}}{\Omega_{\rm m}\gamma} = 0.113 \pm 0.005, \tag{9.37}$$

where the subscript m in our notation corresponds to GRAV in Figure 9.4. Here

$$\gamma = 1 + 0.19\sqrt{h}$$

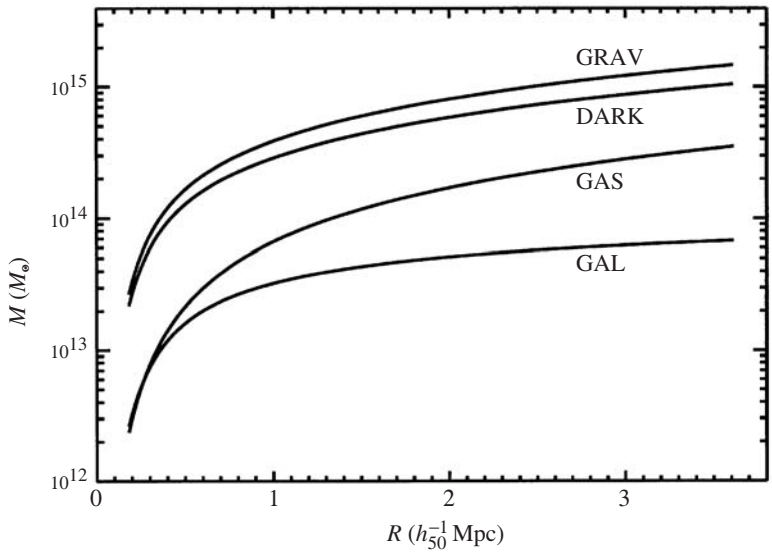


Figure 9.4 The mass inside given radii of different mass components of the A85 galaxy cluster. GAL, mass in galaxies; GAS, mass in intracluster gas; GRAV, total gravitating mass; DARK, missing mass component. From reference [9, 10] courtesy of J. Nevalainen.

denotes the local enhancement of baryon density in a cluster compared with the universal baryon density, obtained from simulations, and the f_{GAS} value is an average for six clusters studied by the Chandra observatory [11]. Using the Ω_{b} value from our inventory above, one finds

$$\Omega_{\mathrm{m}} \approx 0.34. \tag{9.38}$$

Although not a precision determination of Ω_{m}, it is a demonstration that the missing DM in galaxy clusters is considerable.

The amount of DM in rich clusters can also be estimated from the cosmic shear in weak lensing. The advantage of this method is that it does not depend on the radiation emitted by matter. It gives a value of $\Omega_{\mathrm{m}} \approx 0.3$, but not yet with an interesting precision.

Strong evidence for DM comes from the simulation structures and comparison with the structures in the sky. We shall come to this subject in Section 9.5.

9.4 Dark Matter Candidates

If only a few per cent of the total mass of the Universe is accounted for by stars and hydrogen clouds, could baryonic matter in other forms make up DM? The answer given by nucleosynthesis is a qualified no: all baryonic DM is already included in Ω_{b}.

Dark Baryonic Matter. Before the value of $\Omega_{\rm dm}$ was pinned down as certainly as it is now, several forms of dark baryonic matter was considered. Gas or dust clouds were the first thing that came to mind. We have already accounted for hot gas because it is radiating and therefore visible. Clouds of cold gas would be dark but they would not stay cold forever. Unless there exists vastly more cold gas than hot gas, which seems unreasonable, this DM candidate is insufficient.

It is known that starlight is sometimes obscured by *dust*, which in itself is invisible if it is cold and does not radiate. However, dust grains re-radiate starlight in the infrared, so they do leave a trace of their existence. But the amount of dust and rocks needed as DM would be so vast that it would have affected the composition of the stars. For instance, it would have prevented the formation of low-metallicity (population-II) stars. Thus dust is not an acceptable candidate.

Snowballs of frozen hydrogenic matter, typical of comets, have also been considered, but they would sublimate with time and become gas clouds. A similar fate excludes *collapsed stars*: they eject gas which would be detectable if their number density were sufficient for DM.

A more serious candidate for baryonic matter has been *jupiters* or *brown dwarfs*: stars of mass less than $0.08M_\odot$. They also go under the acronym MACHO for *Massive Compact Halo Object*. They lack sufficient pressure to start hydrogen burning, so their only source of luminous energy is the gravitational energy lost during slow contraction. Such stars would clearly be very difficult to see since they do not radiate. However, if a MACHO passes exactly in front of a distant star, the MACHO would act as a gravitational microlens, because light from the star bends around the massive object. The intensity of starlight would then be momentarily amplified (on a timescale of weeks or a few months) by microlensing, as described in Section 2.6. The problem is that even if MACHOs were relatively common, one has to monitor millions of stars for one positive piece of evidence. At the time of writing only a few microlensing MACHOs have been discovered in the space between Earth and the Large Magellanic Cloud [12, 13], but their contribution to $\Omega_{\rm B}$ cannot be precisely evaluated.

The shocking conclusion is that the predominant form of matter in the Universe is nonbaryonic, and we do not even know what it is composed of! Thus we are ourselves made of some minor pollutant, a discovery which may well be called the fourth breakdown of the anthropocentric view. The first three were already accounted for in Chapter 1.

Black Holes. Primordial black holes could be good candidates because they evade the nucleosynthesis bound, they are not luminous, they (almost) do not radiate, and if they are big enough they have a long lifetime, as we saw in Equation (3.30). They are believed to sit at the centre of every galaxy and have masses exceeding $100M_\odot$. The mass range 0.3–$30M_\odot$ is excluded by the nonobservation of MACHOs in the galactic halo (cf. Section 3.4). Various astrophysical considerations limit their mass to around $10^4M_\odot$. But black holes do not appear to be a solution to the galactic rotation curves which require radially distributed DM in the haloes.

CDM. Particles which were very slow at time $t_{\rm eq}$ when galaxy formation started are candidates for CDM. If these particles are massive and have weak interactions, so called WIMPs (*Weakly Interacting Massive Particles*), they became nonrelativistic much earlier than the leptons and become decoupled from the hot plasma. For instance, the supersymmetric models briefly discussed in Section 6.6 contain a very large number of particles, of which the lightest ones would be stable. At least three such neutral SUSY 'sparticles'—the *photino*, the *Zino* and the *Higgsino*—or a linear combination of them (the *neutralino*) could serve. Laboratory searches have not found them up to a mass of about 37 GeV, but they could be as heavy as 1 TeV.

Very heavy neutrinos, $m_\nu > 45$ GeV, could also be CDM candidates, other cold thermal relics of mass up to 300 TeV, and superheavy nonthermal *wimpzillas* at inflaton scale with masses of 10^9–10^{19} GeV. The latter are produced by the gravitational expansion and could also be useful for explaining extremely hard cosmic γ-rays.

Alternatively, the CDM particles may be very light if they have some superweak interactions, in which case they froze out early when their interaction rate became smaller than the expansion rate, or they never even attained thermal equilibrium. Candidates in this category are the *axion* and its SUSY partner *axino*. The axion is a light pseudoscalar boson with a 2γ coupling like the π^0, so it could convert to a real photon by exchanging a virtual photon with a proton. Its mass is expected to be of the order of 1 μeV to 10 meV. It was invented to prevent CP violation in QCD, and it is related to a slightly broken baryon number symmetry in a five-dimensional space-time. Another CDM candidate could be axion clusters with masses of the order of $10^{-8} M_\odot$.

Among further exotica are *solitons*, which are nontopological scalar-field quanta with conserved global charge Q (*Q-balls*) or baryonic charge B (*B-balls*).

The WIMPs would traverse terrestrial particle detectors with a typical virial velocity of the order of 200 km s^{-1}, and perhaps leave measurable recoil energies in their elastic scattering with protons. The proof that the recoil detected was due to a particle in the galactic halo would be the annual modulation of the signal. Because of the motion of the Earth around the Sun, the signal should have a maximum in June and a minimum in December. Several experiments to detect such signals are currently running or being planned, but so far the absence of signals only permits us to set upper limits to the WIMP flux.

All WIMPs have in common that they are hitherto unobserved particles which only exist in some theories. A signal worth looking for would be monoenergetic photons from their annihilation

$$\mathrm{X}_{\rm dm} + \bar{\mathrm{X}}_{\rm dm} \longrightarrow 2\gamma. \tag{9.39}$$

Several experiments are planned or under way to observe these photons if they exist.

WIMP Distribution. The ideal fluid approximation which is true for the collisionless WIMPs on large scales breaks down when they decouple from the

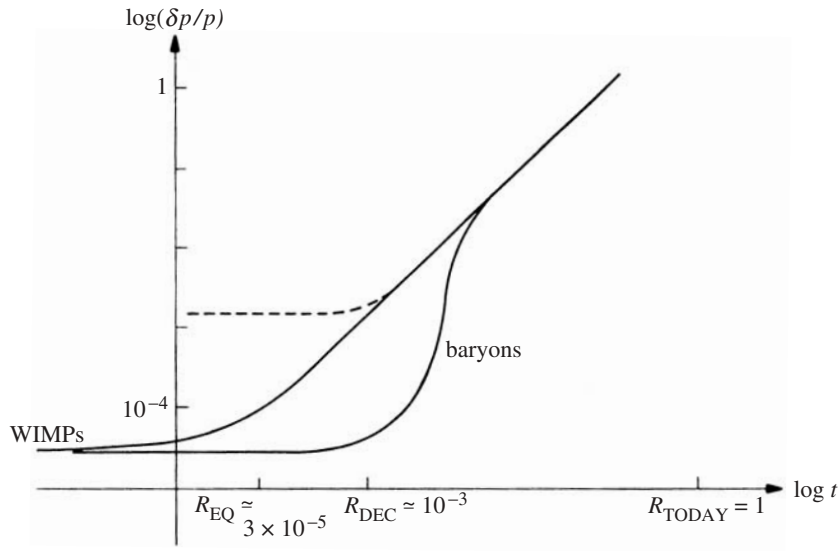


Figure 9.5 The evolution of density fluctuations $\delta\rho/\rho$ in the baryon and WIMP components. The perturbations in the WIMPs begin to grow at the epoch of matter–radiation equality. However, the perturbations in the baryons cannot begin to grow until just after decoupling, when baryons fall into the WIMP potential wells. Within a few expansion times the baryon perturbations 'catch up' with the WIMP perturbations. The dashed line shows where the baryonic density fluctuations would have to start if DM were purely baryonic. Courtesy of E. W. Kolb and M. S. Turner.

plasma and start to stream freely out of overdense regions and into underdense regions, thereby erasing all small inhomogeneities (*Landau damping*). This defines the characteristic length and mass scales for freely streaming particles of mass $m_{\rm dm}$,

$$\lambda_{\rm fs} \simeq 40\left(\frac{30\ {\rm eV}}{m_{\rm dm}}\right)\ {\rm Mpc}, \tag{9.40}$$

$$M_{\rm fs} \simeq 3\times 10^{15}\left(\frac{30\ {\rm eV}}{m_{\rm dm}}\right)^2 M_\odot. \tag{9.41}$$

Perturbations in CDM start growing from the time of matter–radiation equality, while baryonic fluctuations are inhibited until recombination because of the tight coupling with photons (or alternatively one can say because of the large baryonic Jeans mass prior to recombination). After recombination, the baryons fall into the CDM potential wells. A few expansion times later, the baryon perturbations catch up with the WIMPs, and both then grow together until $\delta > 1$, when perturbations become Jeans unstable, collapse and virialize. The amplitude of radiation, however, is unaffected by this growth, so the CMB anisotropies remain at the level determined by the baryonic fluctuations just before recombination. This is illustrated in Figure 9.5.

The lightest WIMPs are slow enough at time $t_{\rm eq}$ to be bound in perturbations on galactic scales. They should then be found today in galaxy haloes together

with possible baryonic MACHOs. If the nonbaryonic DM in our galactic halo were to be constituted by WIMPs at a sufficient density, they should also be found inside the Sun, where they lose energy in elastic collisions with the protons, and ultimately get captured. They could then contribute to the energy transport in the Sun, modifying solar dynamics noticeably. So far this possible effect has not been observed.

On the other hand, if the WIMP overdensities only constituted early potential wells for the baryons, but did not cluster so strongly, most WIMPs would have leaked out now into the intergalactic space. In that case the WIMP distribution in clusters would be more uniform than the galaxy (or light) distribution, so that galaxies would not trace mass.

Hot and Warm Dark Matter. Although the neutrinos decoupled from the thermal plasma long before matter domination, they remained relativistic for a long time because of their small mass. For this reason they would possibly constitute *hot dark matter* (HDM), freely streaming at $t_{\rm eq}$. The accretion of neutrinos to form haloes around the baryon clumps would be a much later process. The CMB is then very little perturbed by the clumps, because most of the energy is in neutrinos and in radiation. However, we already know that the neutrino fraction is much too small to make it a DM candidate, so HDM is no longer a viable alternative.

An intermediate category is constituted by possible sterile neutrinos and by the *gravitino*, which is a SUSY partner of the graviton. These have been called *warm dark matter* (WDM). Both HDM and WDM are now ruled out by computer simulations of the galaxy distribution in the sky. WDM is also ruled out by the WMAP detection of early re-ionization at $z > 10$ [1]. We shall therefore not discuss these alternatives further.

9.5 The Cold Dark Matter Paradigm

The ΛCDM paradigm (which could simply be called CDM, since CDM without a Ω_λ component is ruled out) is based on all the knowledge we have assembled so far: the FLRW model with a spatially flat geometry, BBN and thermodynamics with a known matter inventory including dark energy of unknown origin but known density; inflation-caused linear, adiabatic, Gaussian mass fluctuations accompanying the CMB anisotropies with a nearly scale-invariant Harrison–Zel'dovich power spectrum; growth by gravitational instability from $t_{\rm eq}$ until recombination, and from hot gas to star formation and hierarchical clustering.

The new element in this scenario is collisionless DM, which caused matter domination to start much earlier than if there had been only baryons. The behaviour of DM is governed exclusively by gravity (unless we discover any DM interactions with matter or with itself), whereas the formation of the visible parts of galaxies involves gas dynamics and radiative processes.

While the CMB temperature and polarization anisotropies measure fluctuations at recombination, the galaxy distribution measures fluctuations up to present times. Cosmic shear in weak lensing is sensitive to the distribution of DM directly, but it leaves a much weaker signal than do clusters.

Hierarchical Scenarios. Early CDM models (without an Ω_λ component) produced galaxies naturally, but underproduced galaxy clusters and supergalaxies of mass scale $10^{15} M_\odot$. This was an example of a bottom–top scenario, where small-scale structures were produced first and large-scale structures had to be assembled from them later. Although there was not time enough in this scenario to produce large-scale structures within the known age of the Universe, the scenario could be improved by the introduction of a new degree of freedom, the cosmological constant.

The opposite 'top–bottom' scenario was predicted by HDM models where the first structures, supergalaxies, formed by neutrino clouds contracting into pancakes which subsequently collapsed and disintegrated. Smaller structures and galaxies formed later from the crumbs. But computer simulations of pancake formation and collapse show that the matter at the end of the collapse is so shocked and so heated that the clouds do not condense but remain ionized, unable to form galaxies and attract neutrino haloes. Moreover, large clusters (up to $10^{14} M_\odot$) have higher escape velocities, so they should trap five times more neutrinos than large galaxies of size $10^{12} M_\odot$. This scenario is not supported by observations, which show that the ratio of dynamic mass to luminous mass is about the same in objects of all sizes.

The 'bottom–top' scenario is supported by the observations that supergalaxies are typically at distances $z \lesssim 0.5$, whereas the oldest objects known are quasars at redshifts up to $z = 5$–7. There are also several examples of galaxies which are older than the groups in which they are now found. Moreover, in our neighbourhood the galaxies are generally falling in towards the Virgo cluster rather than streaming away from it.

Several pieces of evidence indicate that luminous galaxies could have been assembled from the merging of smaller star-forming systems before $z \approx 1$. The Hubble Space Telescope as well as ground-based telescopes have discovered vast numbers of faint blue galaxies at $1 \leqslant z \leqslant 3.5$, which obviously are very young. There is also evidence that the galaxy merger rate was higher in the past, increasing roughly as a^{-m} or $(1 + z)^m$ with $m \approx 2$–3. All this speaks for a bottom–top scenario.

Galaxy Surveys. Two large ongoing surveys of the nearby Universe are the two-degree Field Galaxy Redshift Survey (2dFGRS), which has reported studies on nearly 250 000 galaxies within $z < 0.25$ (blue magnitude limit 19.45) [2, 14, 15], and the Sloan Digital Sky Survey (SDSS), which will provide data on a million galaxies out to a magnitude of 23 [16]. At the time of writing, only the results of the 2dFGRS studies have been reported.

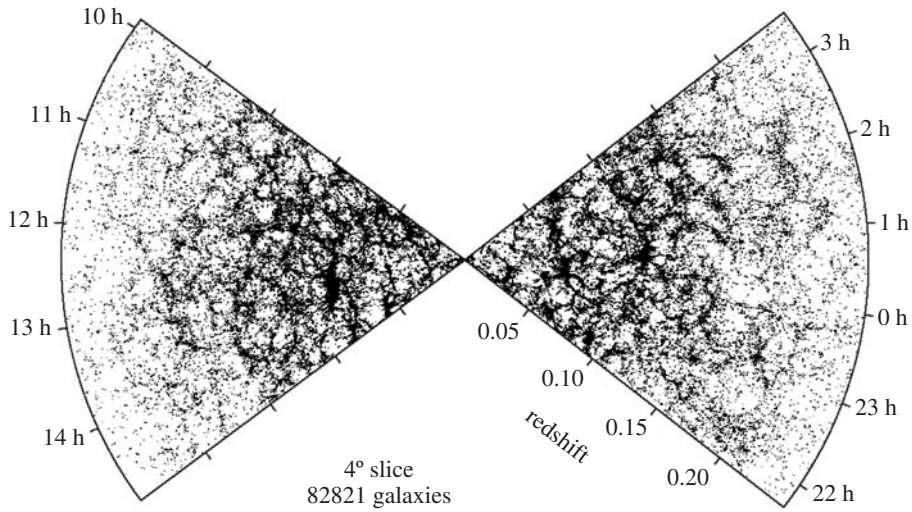


Figure 9.6 The distribution of galaxies in part of the 2dFGRS, drawn from a total of 213 703 galaxies. Reproduced from reference [14] by permission of the 2dFGRS Team.

The purpose of 2dFGRS is, to cite J. A. Peacock [14],

(i) 'To measure the galaxy power spectrum $P(k)$ on scales up to a few hundred Mpc, bridging the gap between the scales of nonlinear structures and measurements from the CMB';

(ii) 'To measure the redshift-space distortion of the large-scale clustering that results from the peculiar velocity field produced by the mass distributions'; and

(iii) 'To measure higher-order clustering statistics in order to understand biased galaxy formation, and to test whether the galaxy distribution on large scales is a Gaussian random field'.

Distributions of galaxies in two-dimensional pictures of the sky show that they form long filaments separating large underdense voids with diameters up to $60h^{-1}$ Mpc. Figure 9.6 shows such a picture of 82 821 galaxies from a total of 213 703 galaxies selected by 2dFGRS. The image reveals a wealth of detail, including linear supercluster features, often nearly perpendicular to the line of sight.

In three-dimensional pictures of the Universe seen by the earlier Infrared Astronomical Satellite IRAS [17, 18] in an all-sky redshift survey, the filaments turn out to form dense sheets of galaxies, of which the largest one is the 'Great Wall', which extends across $170h^{-1}$ Mpc length and $60h^{-1}$ Mpc width.

Large Scale Structure Simulation. The formation and evolution of cosmic structures is so complex and nonlinear and the number of galaxies considered so enormous that the theoretical approach must make use of either numerical simulations or semi-analytic modelling. The strategy in both cases is to calculate how

density perturbations emerging from the Big Bang turn into visible galaxies. As summarized by C. S. Frenk, J. A. Peacock and collaborators in the 2dFGRS Team [2], this requires a number of processes in a phenomenological manner:

(i) the growth of DM haloes by accretion and mergers;

(ii) the dynamics of cooling gas;

(iii) the transformation of cold gas into stars;

(iv) the spectrophotometric evolution of the resulting stellar populations;

(v) the feedback from star formation and evolution on the properties of prestellar gas; and

(vi) the build-up of large galaxies by mergers.

The primary observational information consists of a count of galaxy pairs in the redshift space (σ, π). From this, the correlation function $\xi(s)$ in redshift space, and subsequently the correlation function $\xi(r)$ in real space, can be evaluated. Recall that $\xi(s)$ and $\xi(r)$ are related by Equation (9.33) via the parameter β in Equations (9.32). From $\xi(r)$, the power spectrum $P(k)$ can in principle be constructed using its definition in Equations (9.8) and (9.9).

The observed count of galaxy pairs is compared with the count estimated from a randomly generated mass distribution following the same selection function both on the sky and in redshift. Different theoretical models generate different simulations, depending on the values of a large number of adjustable parameters: h, $\Omega_{\rm m}h \equiv (\Omega_{\rm dm}h^2 + \Omega_{\rm b}h^2)/h$, $\Omega_{\rm b}/\Omega_{\rm m}$, Ω_0, $n_{\rm s}$, the normalization σ_8 and the bias b between galaxies and mass.

The CDM paradigm sets well-defined criteria on the real fluctuation spectrum. A good fit then results in parameter values. Since the parameter combinations here are not the same as in the CMB analysis, the degeneracy in the 2dFGRS data between $\Omega_{\rm m}h$ and $\Omega_{\rm b}/\Omega_{\rm m}$ can be removed by combining the CMB and 2dFGRS analyses. Let us now summarize a few of the results.

If the simulated mass-correlation function $\xi_{\rm dm}(r)$ and the observed galaxy-number two-point-correlation function $\xi_{\rm gal}(r)$ are identical, this implies that light (from galaxies) traces mass exactly. If not, they are biased to a degree described by b. The result is that there is no bias at large scales, as indeed predicted by theory, but on small scales some anti-bias is observed. This result is a genuine success of the theory because it does not depend on any parameter adjustments. Independently, weak lensing observations also show that visible light in clusters does trace mass (all the visible light is emitted by the stars in galaxies, not by diffuse emission), but it is not clear whether this is true on galaxy scales.

Theoretical models predict that the brightest galaxies at $z = 3$ should be strongly clustered, which is indeed the case. This comparison is also independent of any parameter adjustments. In contrast, DM is much more weakly clustered at $z = 3$ than at $z = 0$, indicating that galaxies were strongly biased at birth.

In Figure 9.7 we show the theoretical linear power spectrum $P(k)$. The real 2dFGRS galaxy power spectrum data lie so accurately on the theoretical curve in

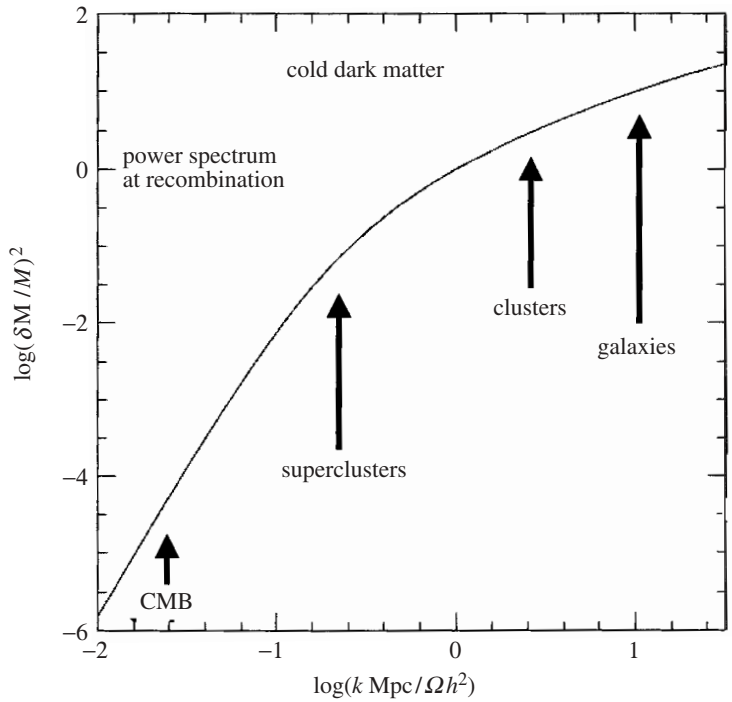


Figure 9.7 The power function $P(k)$ as a function of the density fluctuation wavenumber k in units of h^2 Mpc^{-1}. This can also be expressed by the angular scale in degrees or by the linear size L of present cosmic structures in units of Mpc $\Omega_{\rm m}^{-1}h^{-1}$. Courtesy of C. Frenk and J. Peacock

Figure 9.7 that we have refrained from plotting them. To achieve this success, all the free parameters have been adjusted.

One important signature of gravitational instability is that material collapsing around overdense regions should exhibit peculiar velocities and infall leading to redshift-space distortions of the correlation as shown in Figure 9.3. We have previously referred to large-scale bulk flows of matter observed within the LSC attributed to the 'Great Attractor', an overdensity of mass about $5.4 \times 10^{16} M_\odot$ in the direction of the *Hydra–Centaurus* cluster, but far behind it, at a distance of some $44h^{-1}$ Mpc. The 2dFGRS has verified that both types of redshift-space distortions occur, the 'Fingers of God' due to nearby peculiar velocities and the flattening due to infall at larger distances. These results are quantified by the parameter value

$$\beta = \Omega_{\rm m}^{0.6}/b = 0.43 \pm 0.07. \tag{9.42}$$

With a large-scale bias of $b = 1$ to a precision of about 10%, the β error dominates, so that one obtains

$$\Omega_{\rm m} = 0.25 \pm 0.07, \tag{9.43}$$

in good agreement with other determinations.

The WMAP collaboration [1] obtained a value for the characteristic amplitude of velocity fluctuations within 8 Mpc h^{-1} spheres at $z = 0$,

$$\sigma_8 = 0.9 \pm 0.1,$$

by setting $\beta = \sigma_8 \Omega_{\rm m}^{0.6}$.

This result from the 2dFGRS, as well as other parametric results, is very useful and was combined with CMB data in Section 8.4.

To summarize one can state that on scales larger than a few Mpc the distribution of DM in CDM models is essentially understood. Understanding the inner structure of DM haloes and the mechanisms of galaxy formation has proved to be much more difficult.

Problems

1. The mean free path ℓ of photons in homogeneous interstellar dust can be found from Equation (1.4) assuming that the radius of dust grains is 10^{-7} m. Extinction observations indicate that $\ell \approx 1$ kpc at the position of the Solar System in the Galaxy. What is the number density of dust grains [19]?
2. Derive Equation (9.35). On average the square of the mean random velocity v^2 of galaxies in spherical clusters is three times larger than V^2, where V is the mean line-of-sight velocity displacement of a galaxy with respect to the cluster centre. Calculate M for $V = 1000$ km s^{-1} and $R = 1$ Mpc in units of $M_\odot$ [19].
3. Suppose that galaxies have flat rotation curves out to $R_{\rm max}$. The total mass inside $R_{\rm max}$ is given by Equation (9.26), where v may be taken to be 220 km s^{-1}. If the galaxy number density is $n = 0.01h^3/{\rm Mpc}^3$, show that $\Omega = 1$ when $R_{\rm max}$ is extended out to an average intergalactic distance of $2.5h^{-1}$ Mpc [3].
4. To derive Jeans wavelength $\lambda_{\rm J}$ and Jeans mass $M_{\rm J}$ (see Equation (9.22)), let us argue as follows. A mass $M_{\rm J} = \rho\lambda_{\rm J}^3$ composed of a classical perfect gas will collapse gravitationally if its internal pressure $P = \rho kT/m$ cannot withstand the weight of a column of material of unit area and height $\lambda_{\rm J}$. Here m is the mean mass of the particles forming the gas. If we set the weight of the latter greater than or equal to P,

$$\frac{GM_{\rm J}}{\lambda_{\rm J}^2}\rho\lambda_{\rm J} \gtrsim \frac{\rho kT}{m},$$

   we will obtain a constraint on the sizes of fragment which will separate gravitationally out of the general medium. Show that this leads to Equation (9.22) [19].
5. Suppose that neutralinos have a mass of 100 GeV and that they move with a virial velocity of 200 km s^{-1}. How much recoil energy would they impart to a germanium nucleus?

6. The universal luminosity density radiated in the blue waveband by galaxies is
$$L_{\rm U} = (2 \pm 0.2) \times 10^8 h\, L_\odot\ {\rm Mpc}^{-3}.$$
Show that the Coma value $Y = M/L = 300$ in solar units then gives $\Omega_{\rm m} = 0.30$.
7. Assuming that galaxies form as soon as there is space for them, and that their mean radius is $30h^{-1}$ kpc and their present mean number density is $0.03h^3\ {\rm Mpc}^{-3}$, estimate the redshift at the time of their formation [2].

Chapter Bibliography

[1] Bennett, C. L. *et al*. 2003 Preprint arXiv, astro-ph/0302207 and 2003 *Astrophys. J.* (In press.) and companion papers cited therein.
[2] Hawkins, E. *et al*. 2002 arXiv astro-ph/0212375 and *Mon. Not. R. Astron. Soc.* (2003).
[3] Peebles, P. J. E. 1993 *Principles of Physical Cosmology*. Princeton University Press, Princeton, NJ.
[4] Tremaine, S. and Gunn J. E. 1979 *Phys. Rev. Lett.* **42**, 407.
[5] Battaner, E. *et al*. 1992 *Nature* **360**, 65.
[6] Navarro, J. F., Frenk, C. S. and White, S. D. M. 1996 *Astrophys. J.* **462**, 563.
[7] Corbelli, E. 2003 arXiv astro-ph/0302318 and *Mon. Not. R. Astron. Soc.*
[8] Ponman, T. J. and Bertram, D. 1993 *Nature* **363**, 51.
[9] David, L. P., Jones, C. and Forman, W. 1995 *Astrophys. J.* **445**, 578.
[10] Nevalainen, J. *et al*. 1997 *Proc. 2nd Integral Workshop* (ed. C. Winkler *et al*.), European Space Agency Special Publication no. 382.
[11] Allen, S. W., Schmidt, R. W. and Fabian, A. C. 2002 *Mon. Not. R. Astron. Soc.* **334**, L11.
[12] Alcock, C., Akerlof, C.-W., Allsman, R. A. *et al*. 1993 *Nature* **365**, 621.
[13] Auborg, E., Bareyre, P., Bréhin, S. *et al*. 1993 *Nature* **365**, 623.
[14] Peacock, J. A. 2002 In *A new era in cosmology* (ed. T. Shanks and N. Metcalfe). ASP Conference Proceedings Series.
[15] Frenk, C. S. 2002 *Phil. Trans. R. Soc. Lond.* A **360**, 1277.
[16] York, D. G. *et al*. 2000 *Astr. J.* **120**, 1579.
[17] Moore, R. L., Frenk, C. S., Weinberg, D. *et al*. 1992 *Mon. Not. R. Astron. Soc.* **256**, 477.
[18] Saunders, W., Frenk, C. S., Rowan-Robinson, M. *et al*. 1991 *Nature*, **349**, 32.
[19] Shu, F. H. 1982 *The physical universe*. University Science Books, Mill Valley, CA.

10

Epilogue

We have now covered most of cosmology briefly and arrived happily at a Standard Model. However, it has many flaws and leaves many questions unanswered. Of course the biggest question is what caused the Universe? What caused the Big Bang in the first place, what caused it to be followed by inflation, what caused the inflation to stop, and what caused the properties of the elementary particles and their interactions?

In Section 10.1 we discuss the properties of the initial singularity and related singularities in black holes and in a Big Crunch.

In Section 10.2 we learn that there is a difference between The Beginning and The End. This defines a thermodynamically preferred direction of time. We also briefly discuss extra dimensions and some other open questions. We close with a brief outlook into the future of the Universe.

10.1 Singularities

The classical theory of gravity cannot predict how the Universe began. Gravity itself curls up the topology of the manifold on which it acts, and in consequence singularities can appear. The problem at time zero arises because time itself is created at that point. In studies of the global structure of space-time, Stephen Hawking and Roger Penrose have found [1, 2, 3] that gravity predicts singularities in two situations. One is the Big Bang 'white hole' singularity in our past at the beginning of time. In analogy with a black hole, we have defined a white hole as something which only emits particles and radiation, but does not absorb anything. The other case is a future with gravitationally collapsing stars and other massive bodies, which end up as black-hole singularities. In a closed universe, this is the ultimate Big Crunch at the end of time. In a singularity, the field equations of general relativity break down, so one can only say that 'the theory predicts that it cannot predict the Universe' [3]. Since the Universe is, and has been for some time, undergoing an accelerated expansion we shall not study a Big Crunch.

Introduction to Cosmology Third Edition by Matts Roos
© 2003 John Wiley & Sons, Ltd ISBN 0 470 84909 6 (cased) ISBN 0 470 84910 X (pbk)

Black Hole Analogy. Perhaps something can be learned about our beginning and our end by studying the properties of black-hole singularities: the only ones we may have a chance to observe indirectly. As we saw in Section 3.4, the Friedmann–Lemaître equations are singular $t = 0$. On the other hand, the exponential de Sitter solution (4.60), $a(t) = \exp(Ht)$, is regular at $t = 0$. In the Schwartzschild metric (3.21) the coefficient of $\mathrm{d}t^2$ is singular at $r = 0$, whereas the coefficient of $\mathrm{d}r^2$ is singular at $r = r_\mathrm{c}$. However, if we make the transformation from the radial coordinate r to a new coordinate u defined by

$$u^2 = r - r_\mathrm{c},$$

the Schwartzschild metric becomes

$$\mathrm{d}\tau^r = \frac{u^2}{u^2 + r_\mathrm{c}}\,\mathrm{d}t^2 - 4(u^2 + r_\mathrm{c})\,\mathrm{d}u^2.$$

The coefficient of $\mathrm{d}t^2$ is still singular at $u^2 = -r_\mathrm{c}$, which corresponds to $r = 0$, but the coefficient of $\mathrm{d}u^2$ is now regular at $u^2 = 0$.

A similar game can be played with the de Sitter metric (4.63) in which

$$g_{00} = 1 - r^2H^2, \qquad g_{11} = -(1 - r^2H^2)^{-1}.$$

At $r = H^{-1}$, g_{00} vanishes and g_{11} is singular. In the subspace defined by $0 \leqslant r \leqslant H^{-1}$ we can transform the singularity away by the substitution $u^2 = H^{-1} - r$. The new radial coordinate, u, is then in the range $0 \leqslant u \leqslant H^{-1}$ and the nonsingular metric becomes

$$\mathrm{d}s^2 = (1 - r^2H^2)\,\mathrm{d}t^2 - 4H^{-1}(1 + rH)^{-1}\,\mathrm{d}u^2 - r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2). \qquad (10.1)$$

From these examples we see that some singularities may be just the consequence of a badly chosen metric and not a genuine property of the theory. Moreover, singularities often exist only in exact mathematical descriptions of physical phenomena, and not in reality, when one takes into account limitations imposed by observations. Or they do not exist at all, like the North Pole, which is an inconspicuous location on the locally flat surface of a two-sphere. Yet this surface can represent the complex plane, with the infinities $\pm\infty$ and $\pm\mathrm{i}\infty$ located at the North Pole.

Quantum Arguments. Possible escapes from the singularities in gravity could come from considerations outside the classical theory in four-dimensional space-time, for instance going to quantum theory or increasing the dimensionality of space-time. In quantum gravity, the metric g_{ik} is a quantum variable which does not have a precise value: it has a range of possible values and a probability density function over that range. As a consequence, the proper distance from x^i to $x^i + \mathrm{d}x^i$ is also a quantum variable which does not have a precise value; it can take on any value. It has a probability distribution which peaks at the classical expectation value $\langle \mathrm{d}s^2 \rangle$. Or, phrasing this a little more carefully, the quantum variable of proper distance is represented by a state which is a linear combination of all

possible outcomes of an observation of that state. Under the extreme conditions near $t = 0$, nobody is there to observe, so 'observation' of $\langle \mathrm{d}s^2 \rangle$ has to be defined as some kind of interaction occurring at this proper distance.

Similarly, the cosmic scale factor $a(t)$ in classical gravity has the exact limit

$$\lim_{t \to 0} a(t) = 0 \tag{10.2}$$

at the singularity. In contrast, the quantum scale factor does not have a well-defined limit, it fluctuates with a statistical distribution having variance $\langle a^2(t) \rangle$. This expectation value approaches a nonvanishing constant [2]

$$\lim_{t \to 0} \langle a^2(t) \rangle = C > 0. \tag{10.3}$$

This resembles the situation of a proton–electron pair forming a hydrogen atom. Classically the electron would spiral inwards under the influence of the attractive Coulomb potential

$$V = \frac{e^2}{r}, \tag{10.4}$$

which is singular at $r = 0$. In the quantum description of this system, the electron has a lowest stable orbit corresponding to a minimum finite energy, so it never arrives at the singularity. Its radial position is random, with different possible radii having some probability of occurring. Thus only the mean radius is well defined, and it is given by the expectation value of the probability density function.

In fact, it follows from the limit (10.3) that there is a lower bound to the proper distance between two points,

$$\langle \mathrm{d}s^2 \rangle \geqslant \left(\frac{L_\mathrm{P}}{2\pi} \right)^2, \tag{10.5}$$

where L_P is the Planck length 10^{-35} m [2]. The light cone cannot then have a sharp tip at $t = 0$ as in the classical picture. Somehow the tip of the cone must be smeared out so that it avoids the singular point.

James Hartle and Stephen Hawking [4] have studied a particular model of this kind in an empty space-time with a de Sitter metric (10.1) and a cosmological constant. Although this model is unrealistic because our Universe is not empty, it is mathematically manageable and may perhaps lead to new insights. The exponential de Sitter expansion takes place at times $t \gtrsim t_\mathrm{P}$, when the event horizon is $r \lesssim H^{-1}$. As we have just noted above, the coefficient g_{00} in Equation (10.1) vanishes at $r = H^{-1}$. Hartle and Hawking propose that the metric changes smoothly at this point to a purely spatial Euclidean metric, like the one in Equation (2.29), as shown in Figure 10.1. Thus the de Sitter part of space-time contains no 'beginning of time' with value $t = 0$, whereas in the Euclidean hemisphere, where $r > H^{-1}$, there is no time coordinate at all, time has become a spatial coordinate $\tau = \mathrm{i}t$. Thus one could say (if our space-time would permit such a metric) that time emerges gradually from this space without any abrupt coming into being. Time is limited in the past, to within $\langle t^2 \rangle \simeq t_\mathrm{P}^2$, but it has no sharp boundary. In that sense the classical singularity has disappeared, and there is no origin of time. As

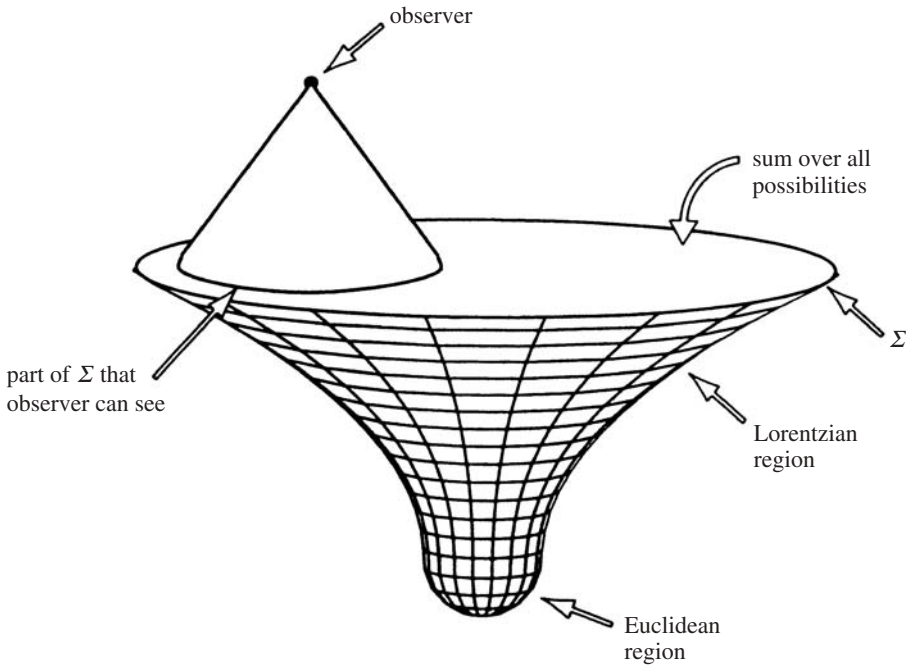


Figure 10.1 An observer can see only part of any surface Σ. From S. Hawking and R. Penrose [3], copyright 1996 by Princeton University Press. Reprinted by permission of Princeton University Press.

Hawking expresses it, 'the boundary condition of the Universe is that it has no boundary' [1, 3].

The Universe then exists because of one unlikely fluctuation of the vacuum, in which the energy ΔE created could be so huge because the time Δt was small enough not to violate Heisenberg's uncertainty relation

$$\Delta t \Delta E \lesssim \hbar. \tag{10.6}$$

However, the step from the Hartle–Hawking universe to a functioning theory of quantum gravity is very large and certainly not mastered.

Some people take the attitude that quantum gravity is unimportant, because inflation has erased all information about what went on before it. Inflation is described to lowest order by classical fields—scalar and perhaps tensor—whereas quantum mechanics only enters as fluctuations.

10.2 Open Questions

The Direction of Time. Since entropy has been increasing all the time and cannot decrease, this defines a preferred direction of time. Therefore the Big Crunch singularity at $t = 2t_{\max}$ in a closed Universe is of the same kind as the singularity in a black hole, and quite unlike the Big Bang white-hole singularity. In classical

gravity, the Universe cannot bounce and turn around to restart a new expansion cycle although classical gravity and quantum mechanics are symmetric in the direction of time, because the second law of thermodynamics is not symmetric. In Section 7.5 we met speculations on a cyclic universe which apparently circumvents this problem.

In his celebrated book *A brief history of time* [1], Stephen Hawking discusses the direction of time, which can be specified in three apparently different ways. Biological time is defined by the ageing of living organisms like ourselves, and by our subjective sense of time (we remember the past but not the future). But this can be shown to be a consequence of the thermodynamical direction of time.

Apparently independent of this is the direction defined by the cosmic expansion which happens to coincide with the thermodynamical arrow of time. Hawking, Laflamme and Lyons [5] have pointed out that the density perturbations also define a direction of time independent of the cosmic time because they always grow, whether the Universe is in an expanding or a contracting phase.

The possible reason why the different arrows of time all point in the same direction is still obscure. Perhaps one has to resort to the Anthropic Principle [6] to 'understand' it, implying that the Universe could not be different if we should be able to exist and observe it. One would then conclude that the agreement between all the different directions of time speaks for the necessity of the Anthropic Principle.

Extra Dimensions. Another enigma is why space-time was created four dimensional. The classical answer was given by *Paul Ehrenfest* (1880–1933), who demonstrated that planetary systems with stable orbits in a central force field are possible only if the number of spatial coordinates is two or three. However, extra space-time dimensions are possible if the higher dimensions are unnoticeably small.

Increasing the dimensionality of space-time to five (or more) may solve many problems naturally. We would then be living on a four-dimensional projection, a *brane*, curved in the fifth dimension at some extremely small *compactification scale*, perhaps the GUT scale, unable to move to any other brane.

The singularity at time zero would only be apparent on our brane, whereas in the full five-dimensional space-time it would correspond to a finite point. Also the *hierarchy problem* could then be solved: the gravitational interaction appears so weak in comparison with the strong and electroweak interactions because the latter are confined to our three-dimensional spatial brane, while gravitation acts in the full five-dimensional space. Gravitation could also be related to supersymmetry in a higher-dimensional space-time.

Different bubbles in Linde's chaotic foam could perhaps be connected through the fifth dimension if different bubbles exist on different branes. Such ideas appear fruitful in many contexts, but they are quite speculative and in addition mathematically complicated, so we shall not treat them further here.

Multiply Connected Universes. We have tacitly assumed that our Universe is simply connected, so that no radial path takes us back to our present space-time location. But this assumption could be abandoned without any dramatic consequences. One can construct many versions of multiply connected universes of finite size, in which one could, in principle, observe our Galaxy at some distant location. Thus the idea is testable.

The problem is, however, that if we search for our Galaxy or some conspicuous pattern of galaxies in another location of the sky, they may look quite different from the way they do now because of the different time perspective and their evolution history. We don't know how the Milky Way looked 10 billion years ago. Searches for conspicuous patterns of quasars or similar galaxies in different directions of the sky have been done, but no convincing evidence has been found.

Fine-Tuning Problems. The origin of the cosmological constant λ is unknown and its value has to be fine-tuned to within the 52nd decimal of zero (in units of $c = 1$). To try to remedy this accident we introduced a time-dependent $\lambda(t)$ initially coupled to inflation which, even so, had to be fine-tuned to within 17 decimal places (Chapter 7, Problem 6). The function $\lambda(t)$ is of course ad hoc, depending on new parameters that characterize the timescale of the deflationary period and the transfer of energy from dark energy to dust.

Tracking quintessence appeared at first sight to remove the need for fine-tuning entirely, but at the cost of the arbitrariness in the choice of a parametric potential $V(\varphi)$. We may perhaps content ourselves with a choice of parameter values fitting the initial conditions and the necessary properties of dark energy, but this does not answer the even deeper coincidence problem of why dark energy has come to dominate visible energy just now, when we are there to observe it.

Dark energy dominates now, its pressure is negative and its equation of state is $w = -1$ or very close to that. Could it cease to dominate later, so that the accelerated expansion ends, perhaps to be replaced by contraction? Is the present the first time in the history of the Universe when dark energy has dominated, or is dark energy a cosmological constant? One can generalize this question and ask whether other constants of nature are indeed constants, or whether it is a coincidence that they have their present values and appear constant.

Missing Physics. The fluctuations during the inflation are the source of the CMB fluctuations at the time of decoupling, but to connect the two eras is difficult. The Universe then traversed the graceful exit with its production of particles and radiation, but even the particle spectrum at that time is unknown to us. Was there an era of supersymmetry, later to be broken?

Equally enigmatic is the nature and origin of DM, of which we only know that it is 'cold', it does not interact with visible matter and it is not identical to dark energy [7]. Does it consist of primordial supersymmetry particles? It is eagerly expected that laboratory particle physics will give the answer.

If we turn to astrophysics at large, there are many unknowns. We haven't

even mentioned the enormously energetic gamma-ray bursts from sources at cosmological distances, the active galactic nuclei (AGN) or the ultra-high energy gamma rays of nearly 10^{21} eV, coming from unknown sources and accelerated by unknown mechanisms. We have discussed many aspects of galaxies, but the astrophysics of galaxy formation is not understood, the thermal history of the intergalactic medium and its baryon and metal content is not known, etc.

The Future. Since the Universe at present undergoes accelerated expansion one would not expect it to halt and turn around towards a Big Crunch. The expansion dilutes the energy density $\rho_{\rm m}$ and the kinetic term of the dark energy, $\frac{1}{2}\dot{\varphi}^2$. If they get much smaller than the potential $V(\varphi)$ in Equations (4.68) or in Equation (7.36), and if it happens that $V(\varphi)$ becomes negative, then the role of the dark energy is inverted: it becomes attractive. The Universe then starts to contract towards a Big Crunch with a rather more easily predictable future.

As contraction proceeds, galaxies start to merge and, when the ambient temperature becomes equal to the interior of the stars, they disintegrate by explosion. Stars are also disrupted by colossal tidal forces in the vicinity of black holes which grow by accreting stellar matter. As the temperature rises, we run the expansion history backwards to recover free electrons, protons and neutrons, subsequently free quarks, and finally free massless GUT fields at time $t_{\rm GUT}$ before the final singularity. The role of black holes is model dependent, but it is reasonable to imagine that all matter ends up in one black hole. What then happens at the singularity escapes all prediction.

In an eternally expanding universe or an asymptotically coasting universe, proton decay is a source of heat for dead stars until about time $\tau_{\rm p}$. Long before $\tau_{\rm p}$, all stars have already exhausted their fuel to become either black holes, neutron stars, black dwarfs or dead planets and their decay products: electrons, positrons, neutrinos and a considerable amount of radiation, all of which are responsible for most of the heat and entropy. The relic CMB has been redshifted away to completely negligible energies. Almost all the energy density of the Universe is dark energy.

From $\tau_{\rm p}$ on, the future is boring [6]. The radiation from decayed protons may cause a brief time of increased radiation, of the order of $1000\tau_{\rm p}$, followed by the redshift of these decay photons to lower enough energies. Then the only thing happening is the very slow formation of positronium by free electrons and positrons. However, these atoms would have little resemblance to present-day positronium. Their size would be of the order of 10^{15} Mpc, and they would rotate around their centre of gravity with an orbital velocity about 1 μm per century. In the end, each of these atoms would decay into some 10^{22} ultrasoft photons at a timescale comparable to the evaporation time of supergalaxy-sized black holes.

Our last speculation concerns the fate of black holes. Since black holes devour matter of all kinds, the outside world will lose all knowledge of what went into the hole. Thus there is a net loss of information or entropy from part of our Universe which is (in principle) observable. Classically, one may argue that the information is not lost, it is just invisible to us inside the event horizon. But

quantum theory permits the black hole to radiate and lose mass without disclosing any other information than the temperature of the radiation. Once a black hole has evaporated completely there will remain no entropy and no information about its contents. Or, there remains a naked singularity, whatever that implies. Then where did the huge amount of information or entropy in the black hole go? Physics does not accept that it just vanished, so this problem has stimulated a flood of speculations.

A singular point cannot simply 'appear' in the middle of space-time in such a way that it becomes 'visible' at some finite future point. We must not be able to observe a particle actually falling into a singularity, where the rules of physics would cease to hold or reach infinity. This is Hawking's and Penrose's hypothesis of cosmic censorship, already referred to in connection with black holes, that singularities should be protected from inspection, either because they exist entirely in the future (Big Crunch), or entirely in the past (Big Bang), or else they are hidden by an event horizon inside black holes [1, 2, 3]. Otherwise, since space-time has its origin in a singularity, perhaps all of space-time would disappear at the appearance of a naked singularity.

Problems

1. Assume that the protons decay with a mean life of $\tau_p = 10^{35}$ yr, converting all their mass into heat. What would the ambient temperature on the surface of Earth be at time $t = \tau_p$, assuming that no other sources of heat contribute?

Chapter Bibliography

[1] Hawking, S. W. 1988 *A brief history of time*. Bantam Books, New York.
[2] Penrose, R. 1989 *The emperor's new mind*. Oxford University Press, Oxford.
[3] Hawking, S. W. and Penrose, R. 1996 *The nature of space and time*. Princeton University Press, Princeton, NJ.
[4] Hartle, J. B. and Hawking, S. W. 1983 *Phys. Rev.* D **28**, 2960.
[5] Hawking, S. W., Laflamme, R. and Lyons, G. W. 1993 *Phys. Rev.* D **47**, 5342.
[6] Barrow, J. D. and Tipler, F. J. 1988 *The Anthropic Cosmological Principle*. Oxford University Press, Oxford.
[7] Sandvik, H. B. *et al*. 2002 arXiv astro-ph/0212114.

Tables

Table A.1 Cosmic distances and dimensions

distance to the Sun	8′ 15″ (light minutes)
distance to the nearest star (α Centauri)	1.3 pc
diameters of globular clusters	5–30 pc
thickness of our Galaxy, the 'Milky Way'	0.3 kpc
distance to our galactic centre	8 kpc
radius of our Galaxy, the 'Milky Way'	12.5 kpc
distance to the nearest galaxy (Large Magellanic Cloud)	55 kpc
distance to the Andromeda nebula (M31)	770 kpc
size of galaxy groups	1–5 Mpc
thickness of filament clusters	$5\ h^{-1}$ Mpc
distance to the Local Supercluster centre (in Virgo)	17 Mpc
distance to the 'Great Attractor'	$44\ h^{-1}$ Mpc
size of superclusters	$\gtrsim 50\ h^{-1}$ Mpc
size of large voids	$60\ h^{-1}$ Mpc
distance to the Coma cluster	$100\ h^{-1}$ Mpc
length of filament clusters	$100\ h^{-1}$ Mpc
size of the 'Great Wall'	$> 60 \times 170 h^{-2}$ Mpc^2
Hubble radius	$3000\ h^{-1}$ Mpc

Introduction to Cosmology Third Edition by Matts Roos
© 2003 John Wiley & Sons, Ltd ISBN 0 470 84909 6 (cased) ISBN 0 470 84910 X (pbk)

Table A.2 Cosmological and astrophysical constants (from the Particle Data Group compilation, K. Hagiwara, *et al*. (2002) *Phys. Rev.* D **66**, 010001-1.)

unit	symbol	value
speed of light	c	299 792 458 m s^{-1}
light year	ly	0.3066 pc = 0.9461×10^{16} m
parsec	pc	3.262 ly = $3.085\,678 \times 10^{16}$ m
solar luminosity	$L_\odot$	$(3.846 \pm 0.008) \times 10^{26}$ J s^{-1}
solar mass	$M_\odot$	1.989×10^{30} kg
solar equatorial radius	$R_\odot$	6.961×10^{8} m
Hubble parameter	H_0	$100h$ km s^{-1} Mpc^{-1} = $h/(9.778\,13$ Gyr)
	h	$0.71^{+0.04}_{-0.03}$
Newtonian constant	G	6.673×10^{-11} m^3 kg^{-1} s^{-2}
Planck constant	$\hbar$	$6.582\,119 \times 10^{-22}$ MeV s
Planck mass	$M_P = \sqrt{\hbar c/G}$	1.221×10^{19} GeV c^2
Planck time	$t_P = \sqrt{\hbar G/c^5}$	5.31×10^{-44} s
Boltzmann constant	k	$8.617\,34 \times 10^{-5}$ eV K^{-1}
Stefan–Boltzmann constant	$a = \pi^2 k^4/15\hbar^3 c^3$	4.7222×10^{-3} MeV m^{-3} K^{-4}
critical density of the Universe	$\rho_C = 3H_0^2/8\pi G$	$2.775 \times 10^{11} h^2 M_\odot$ Mpc^{-3} = $10.538h^2$ GeV m^{-3}

Table A.3 Electromagnetic radiation

type	wavelength [m]	energy [eV]	energy density[1] [eV m^{-3}]
radio	> 1	$<10^{-6}$	≈ 0.05
microwave	1–5×10^{-3}	10^{-6}–2×10^{-3}	3×10^5
infrared	2×10^{-3}–7×10^{-7}	10^{-3}–1.8	?
optical	$(7$–$4) \times 10^{-7}$	1.8–3.1	$\approx 2 \times 10^3$
ultraviolet	4×10^{-7}–10^{-8}	3.1–100	?
X-rays	10^{-8}–10^{-12}	100–10^6	75
γ-rays	$< 10^{-12}$	$> 10^6$	25

[1]From M. S. Longair 1995 In *The Deep Universe* (ed. A. R. Sandage, R. G. Kron, and M. S. Longair), pp. 317. Springer.

Table A.4 Particle masses[1]

particle	MeV units	K units
γ	0	0
$\langle \nu_i \rangle$	$< 0.23 \times 10^{-6}$	$< 2.7 \times 10^{3}$
$e^{\pm}$	0.511	5.93×10^{9}
$\mu^{\pm}$	105.658	1.226×10^{12}
π^0	134.977	1.566×10^{12}
$\pi^{\pm}$	139.570	1.620×10^{12}
p	938.272	1.089×10^{13}
n	939.565	1.090×10^{13}
$\tau^{\pm}$	1 777	2.062×10^{13}
$W^{\pm}$	80 423	9.333×10^{14}
Z	91 188	1.058×10^{15}
H^0	$> 114\,300$	$> 1.326 \times 10^{15}$

[1]From the Particle Data Group compilation K. Hagiwara *et al*. 2002, *Phys. Rev.* D **66**, 010001.

Table A.5 Particle degrees of freedom in the ultrarelativistic limit

particle	particle type	$n_{\rm spin}$	$n_{\rm anti}$	g
γ	vector boson	2	1	2
ν_e, ν_μ, ν_τ	fermion (lepton)	1[1]	2[2]	$\frac{7}{4}$
e^-, μ^-, τ^-	fermion (lepton)	2	2	$\frac{7}{2}$
$\pi^{\pm}$, π^0	boson (meson)	1	1	1
p, n	fermion (baryon)	2	2	$\frac{7}{2}$
$W^{\pm}$, Z	vector boson	3	1	3

[1] $n_{\rm spin} = 2$, but the right-handed neutrinos are inert below the electroweak symmetry breaking.
[2] $n_{\rm anti} = 1$ if the neutrinos are their own antiparticles.

Table A.6 Present properties of the Universe

unit	symbol	value
age	t_0	(13.7 ± 0.2) Gyr
mass	M_U	$\approx 10^{22} M_\odot$
CMB radiation temperature	T_0	2.725 ± 0.001 K
cosmic neutrino temperature	T_ν	1.949 ± 0.001 K
radiation energy density	$\varepsilon_{r,0}$	2.604×10^5 eV m^{-3}
radiation density parameter	Ω_r	$2.471 h^{-2} \times 10^{-5}$
entropy density	s	2.890×10^9 m^{-3}
CMB photon number density	N_γ	4.11×10^8 photons m^{-3}
cosmological constant	$\lvert\lambda\rvert$	$1.3 \times 10^{-52} c^2$ m^{-2}
Schwarzschild radius	$r_{c,\text{Universe}}$	$\gtrsim 11$ Gpc
Baryon to photon ratio	η	$(6.1 \pm 0.7) \times 10^{-10}$
total density parameter	Ω_0	1.02 ± 0.02
baryon density parameter (for $\Omega_0 = 1$)	Ω_b	0.044 ± 0.004
matter density parameter (for $\Omega_0 = 1$)	Ω_m	0.27 ± 0.04
deceleration parameter	q_0	-0.60 ± 0.02

Table A.7 Net baryon number change ΔB and branching fraction BR for leptoquark X decays

i	channel i	ΔB_i	BR_i
1	$\text{X} \to \bar{\text{u}} + \bar{\text{u}}$	$-\frac{2}{3}$	r
2	$\text{X} \to \text{e}^- + \text{d}$	$+\frac{1}{3}$	$1 - r$
3	$\bar{\text{X}} \to \text{u} + \text{u}$	$+\frac{2}{3}$	$\bar{r}$
4	$\bar{\text{X}} \to \text{e}^+ + \text{d}$	$-\frac{1}{3}$	$1 - \bar{r}$

Index

Introduction to Cosmology Third Edition by Matts Roos

© 2003 John Wiley & Sons, Ltd

ISBN 0 470 84909 6 (cased) ISBN 0 470 84910 X (pbk)