An Introduction to Mathematical Cosmology

Second edition

J. N. Islam

CAMBRIDGE

CAMBRIDGE more information - www.cambridge.org/0521496500

This page intentionally left blank

AN INTRODUCTION TO MATHEMATICAL COSMOLOGY

This book provides a concise introduction to the mathematical aspects of the origin, structure and evolution of the universe. The book begins with a brief overview of observational and theoretical cosmology, along with a short introduction to general relativity. It then goes on to discuss Friedmann models, the Hubble constant and deceleration parameter, singularities, the early universe, inflation, quantum cosmology and the distant future of the universe. This new edition contains a rigorous derivation of the Robertson–Walker metric. It also discusses the limits to the parameter space through various theoretical and observational constraints, and presents a new inflationary solution for a sixth degree potential.

This book is suitable as a textbook for advanced undergraduates and beginning graduate students. It will also be of interest to cosmologists, astrophysicists, applied mathematicians and mathematical physicists.

JAMAL NAZRUL ISLAM received his PhD and ScD from the University of Cambridge. In 1984 he became Professor of Mathematics at the University of Chittagong, Bangladesh, and is currently Director of the Research Centre for Mathematical and Physical Sciences, University of Chittagong. Professor Islam has held research positions in university departments and institutes throughout the world, and has published numerous papers on quantum field theory, general relativity and cosmology. He has also written and contributed to several books.

AN INTRODUCTION TO MATHEMATICAL COSMOLOGY

Second edition

J. N. ISLAM

Research Centre for Mathematical and Physical Sciences,
University of Chittagong, Bangladesh



CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

© Cambridge University Press 1992, 2004

First published in printed format 2001

ISBN 0-511-01849-5 eBook (netLibrary)
ISBN 0-521-49650-0 hardback
ISBN 0-521-49973-9 paperback

Contents

Preface to the first edition *page* ix
Preface to the second edition xi
1 Some basic concepts and an overview of cosmology 1
2 Introduction to general relativity 12
2.1 Summary of general relativity 12
2.2 Some special topics in general relativity 18
2.2.1 Killing vectors 18
2.2.2 Tensor densities 21
2.2.3 Gauss and Stokes theorems 24
2.2.4 The action principle for gravitation 28
2.2.5 Some further topics 32
3 The Robertson–Walker metric 37
3.1 A simple derivation of the Robertson–Walker metric 37
3.2 Some geometric properties of the Robertson–Walker metric 42
3.3 Some kinematic properties of the Robertson–Walker metric 45
3.4 The Einstein equations for the Robertson–Walker metric 51
3.5 Rigorous derivation of the Robertson–Walker metric 53
4 The Friedmann models 60
4.1 Introduction 60
4.2 Exact solution for zero pressure 64
4.3 Solution for pure radiation 67
4.4 Behaviour near $t = 0$ 68
4.5 Exact solution connecting radiation and matter eras 68

4.6 The red-shift versus distance relation 71
4.7 Particle and event horizons 73
5 The Hubble constant and the deceleration parameter 76
5.1 Introduction 76
5.2 Measurement of H_0 77
5.3 Measurement of q_0 80
5.4 Further remarks about observational cosmology 85
Appendix to Chapter 5 90
6 Models with a cosmological constant 94
6.1 Introduction 94
6.2 Further remarks about the cosmological constant 98
6.3 Limits on the cosmological constant 100
6.4 Some recent developments regarding the cosmological constant and related matters 102
6.4.1 Introduction 102
6.4.2 An exact solution with cosmological constant 104
6.4.3 Restriction of parameter space 107
7 Singularities in cosmology 112
7.1 Introduction 112
7.2 Homogeneous cosmologies 113
7.3 Some results of general relativistic hydrodynamics 115
7.4 Definition of singularities 118
7.5 An example of a singularity theorem 120
7.6 An anisotropic model 121
7.7 The oscillatory approach to singularities 122
7.8 A singularity-free universe? 126
8 The early universe 128
8.1 Introduction 128
8.2 The very early universe 135
8.3 Equations in the early universe 142
8.4 Black-body radiation and the temperature of the early universe 143
8.5 Evolution of the mass-energy density 148
8.6 Nucleosynthesis in the early universe 153
8.7 Further remarks about helium and deuterium 159
8.8 Neutrino types and masses 164

9 The very early universe and inflation 166
9.1 Introduction 166
9.2 Inflationary models – qualitative discussion 167
9.3 Inflationary models – quantitative description 174
9.4 An exact inflationary solution 178
9.5 Further remarks on inflation 180
9.6 More inflationary solutions 183
Appendix to Chapter 9 186
10 Quantum cosmology 189
10.1 Introduction 189
10.2 Hamiltonian formalism 191
10.3 The Schrödinger functional equation for a scalar field 195
10.4 A functional differential equation 197
10.5 Solution for a scalar field 199
10.6 The free electromagnetic field 199
10.7 The Wheeler–De Witt equation 201
10.8 Path integrals 202
10.9 Conformal fluctuations 206
10.10 Further remarks about quantum cosmology 209
11 The distant future of the universe 211
11.1 Introduction 211
11.2 Three ways for a star to die 211
11.3 Galactic and supergalactic black holes 213
11.4 Black-hole evaporation 215
11.5 Slow and subtle changes 216
11.6 A collapsing universe 218
Appendix 220
Bibliography 238
Index 247

Preface to the first edition

Ever since I wrote my semi-popular book *The Ultimate Fate of the Universe* I have been meaning to write a technical version of it. There are of course many good books on cosmology and it seemed doubtful to me whether the inclusion of a chapter on the distant future of the universe would itself justify another book. However, in recent years there have been two interesting developments in cosmology, namely inflationary models and quantum cosmology, with their connection with particle physics and quantum mechanics, and I believe the time is ripe for a book containing these topics. Accordingly, this book has a chapter each on inflationary models, quantum cosmology and the distant future of the universe (as well as a chapter on singularities not usually contained in the standard texts).

This is essentially an introductory book. None of the topics dealt with have been treated exhaustively. However, I have tried to include enough introductory material and references so that the reader can pursue the topic of his interest further.

A knowledge of general relativity is helpful; I have included a brief exposition of it in Chapter 2 for those who are not familiar with it. This material is very standard; the form given here is taken essentially from my book *Rotating Fields in General Relativity*.

In the process of writing this book, I discovered two exact cosmological solutions, one connecting radiation and matter dominated eras and the other representing an inflationary model for a sixth degree potential. These have been included in Sections 4.5 and 9.4 respectively as I believe they are new and have some physical relevance.

I am grateful to J. V. Narlikar and M. J. Rees for providing some useful references. I am indebted to a Cambridge University Press reader for helpful comments; the portion on observational cosmology has I believe improved considerably as a result of these comments. I am grateful to

F. J. Dyson for his ideas included in the last chapter. I thank Maureen Storey of Cambridge University Press for her efficient and constructive subediting.

I am grateful to my wife Suraiya and daughters Nargis and Sadaf and my son-in-law Kamel for support and encouragement during the period this book was written. I have discussed plans for my books with Mrs Mary Wraith, who kindly typed the manuscript for my first book. For more than three decades she has been friend, philosopher and mentor for me and my wife and in recent years a very affectionate godmother ('Goddy') to my daughters. This book is fondly dedicated to this remarkable person.

Jamal Nazrul Islam
Chittagong, 1991

Preface to the second edition

The material in the earlier edition, to which there appears to have been a favourable response, has been kept intact as far as possible in this new edition except for minor changes. A number of new additions have been made. Some standard topics have been added to the introduction to general relativity, such as Killing vectors. Not all these topics are used later in the book, but some may be of use to the beginning student for mathematical aspects of cosmological studies. Observational aspects have been brought up to date in an extended chapter on the cosmological constant. As this is a book on mathematical cosmology, the treatment of observations is not definitive or exhaustive by any means, but hopefully it is adequate. To clarify the role of the cosmological constant, much discussed in recent years, an exact, somewhat unusual solution with cosmological constant is included. Whether the solution is new is not clear: it is meant to provide a 'comprehension exercise'. One reviewer of the earlier edition wondered why the Hubble constant and the deceleration parameter were chosen for a separate chapter. I believe these two parameters are among the most important in cosmology; adequate understanding of these helps to assess observations generally. Within the last year or two, through analyses of supernovae in distant galaxies, evidence seems to be emerging that the universe may be accelerating, or at least the deceleration may be not as much as was supposed earlier. If indeed the universe is accelerating, the nomenclature 'deceleration parameter' may be called into question. In any case, much more work has to be done, both observational and theoretical, to clarify the situation and it is probably better to retain the term, and refer to a possible acceleration as due to a 'negative deceleration parameter' (in case one has to revert back to 'deceleration'!). I believe it makes sense, in most if not all subjects, constantly to refer back to earlier work, observational, experimental or practical, as well as theoretical aspects, for

this helps to point to new directions and to assess new developments. Some of the material retained from the first edition could be viewed in this way.

A new exact inflationary solution for a sixth degree potential has been added to the chapter on the very early universe. The chapter on quantum cosmology is extended to include a discussion on functional differential equations, material which is not readily available. This topic is relevant for an understanding of the Wheeler–De Witt equation. Some additional topics and comments are considered in the Appendix at the end of the book. Needless to say, in the limited size and scope of the book an exhaustive treatment of any topic is not possible, but we hope enough ground has been covered for the serious student of cosmology to benefit from it.

As this book was going to press, Fred Hoyle passed away. Notwithstanding the controversies he was involved in, I believe Hoyle was one of the greatest contributors to cosmology in the twentieth century. The controversies, more often than not, led to important advances. Hoyle's prediction of a certain energy level of the carbon nucleus, revealed through his studies of nucleosynthesis, confirmed later in the laboratory, was an outstanding scientific achievement. A significant part of my knowledge of cosmology, for what it is worth, was acquired through my association with the then Institute of Theoretical Astronomy at Cambridge, of which the Founder-Director was Hoyle, who was kind enough to give me an appointment for some years. I shall always remember this with gratitude.

I am grateful to Clare Hall, Cambridge, for providing facilities where the manuscript and proofs were completed.

I am grateful for helpful comments by various CUP readers and referees, although it has not been possible to incorporate all their suggestions. I thank the various reviewers of the earlier edition for useful comments. I am grateful to Simon Mitton, Rufus Neal, Adam Black and Tamsin van Essen for cooperation and help at various stages in the preparation of this edition. I thank 'the three women in my life' (Suraiya, Sadaf and Nargis) and my son-in-law Kamel for support and encouragement.

Jamal Nazrul Islam
Chittagong, November 2000

IN MEMORIAM
Mary Wraith (1908–1995)
in affection, admiration and gratitude

1

Some basic concepts and an overview of cosmology

In this chapter we present an elementary discussion of some basic concepts in cosmology. Although the mathematical formalism is essential, some of the main ideas underlying the formalism are simple and it helps to have an intuitive and qualitative notion of these ideas.

Cosmology is the study of the large-scale structure and behaviour of the universe, that is, of the universe taken as a whole. The term 'as a whole' applied to the universe needs a precise definition, which will emerge in the course of this book. It will be sufficient for the present to note that one of the points that has emerged from cosmological studies in the last few decades is that the universe is not simply a random collection of irregularly distributed matter, but it is a single entity, all parts of which are in some sense in unison with all other parts. This, at any rate, is the view taken in the 'standard models' which will be our main concern. We may have to modify these assertions when considering the inflationary models in a later chapter.

When considering the large-scale structure of the universe, the basic constituents can be taken to be galaxies, which are congregations of about 10^{11} stars bound together by their mutual gravitational attraction. Galaxies tend to occur in groups called clusters, each cluster containing anything from a few to a few thousand galaxies. There is some evidence for the existence of clusters of clusters, but not much evidence of clusters of clusters of clusters or higher hierarchies. 'Superclusters' and voids (empty regions) have received much attention (see Chapter 5). Observations indicate that on the average galaxies are spread uniformly throughout the universe at any given time. This means that if we consider a portion of the universe which is large compared to the distance between typical nearest galaxies (this is of the order of a million light years), then the number of galaxies in that portion is roughly the same as the number in another

portion with the same volume at any given time. This proviso 'at any given time' about the uniform distribution of galaxies is important because, as we shall see, the universe is in a dynamic state and so the number of galaxies in any given volume changes with time. The distribution of galaxies also appears to be isotropic about us, that is, it is the same, on the average, in all directions from us. If we make the assumption that we do not occupy a special position amongst the galaxies, we conclude that the distribution of galaxies is isotropic about any galaxy. It can be shown that if the distribution of galaxies is isotropic about every galaxy, then it is necessarily true that galaxies are spread uniformly throughout the universe.

We adopt here a working definition of the universe as the totality of galaxies causally connected to the galaxies that we observe. We assume that observers in the furthest-known galaxies would see distributions of galaxies around them similar to ours, and the furthest galaxies in their field of vision in the opposite direction to us would have similar distributions of galaxies around them, and so on. The totality of galaxies connected in this manner could be defined to be the universe.

E. P. Hubble discovered around 1930 (see, for example, Hubble (1929, 1936)) that the distant galaxies are moving away from us. The velocity of recession follows Hubble's law, according to which the velocity is proportional to distance. This rule is approximate because it does not hold for galaxies which are very near nor for those which are very far, for the following reasons. In addition to the systematic motion of recession every galaxy has a component of random motion. For nearby galaxies this random motion may be comparable to the systematic motion of recession and so nearby galaxies do not obey Hubble's law. The very distant galaxies also show departures from Hubble's law partly because light from the very distant galaxies was emitted billions of years ago and the systematic motion of galaxies in those epochs may have been significantly different from that of the present epoch. In fact by studying the departure from Hubble's law of the very distant galaxies one can get useful information about the overall structure and evolution of the universe, as we shall see.

Hubble discovered the velocity of recession of distant galaxies by studying their red-shifts, which will be described quantitatively later. The red-shift can be caused by other processes than the velocity of recession of the source. For example, if light is emitted by a source in a strong gravitational field and received by an observer in a weak gravitational field, the observer will see a red-shift. However, it seems unlikely that the red-shift of distant galaxies is gravitational in origin; for one thing these red-shifts are rather large for them to be gravitational and, secondly, it is difficult to understand

the systematic increase with faintness on the basis of a gravitational origin. Thus the present consensus is that the red-shift is due to velocity of recession, but an alternative explanation of at least a part of these red-shifts on the basis of either gravitation or some hitherto unknown physical process cannot be completely ruled out.

The universe, as we have seen, appears to be homogeneous and isotropic as far as we can detect. These properties lead us to make an assumption about the model universe that we shall be studying, called the Cosmological Principle. According to this principle the universe is homogeneous everywhere and isotropic about every point in it. This is really an extrapolation from observation. This assumption is very important, and it is remarkable that the universe seems to obey it. This principle asserts what we have mentioned before, that the universe is not a random collection of galaxies, but it is a single entity.

The Cosmological Principle simplifies considerably the study of the large-scale structure of the universe. It implies, amongst other things, that the distance between any two typical galaxies has a universal factor, the same for any pair of galaxies (we will derive this in detail later). Consider any two galaxies A and B which are taking part in the general motion of expansion of the universe. The distance between these galaxies can be written as $f_{AB}R$, where f_{AB} is independent of time and R is a function of time. The constant f_{AB} depends on the galaxies A and B. Similarly, the distance between galaxies C and D is $f_{CD}R$, where the constant f_{CD} depends on the galaxies C and D. Thus if the distance between A and B changes by a certain factor in a definite period of time then the distance between C and D also changes by the same factor in that period of time. The large-scale structure and behaviour of the universe can be described by the single function R of time. One of the major current problems of cosmology is to determine the exact form of $R(t)$. The function $R(t)$ is called the scale factor or the radius of the universe. The latter term is somewhat misleading because, as we shall see, the universe may be infinite in its spatial extent in which case it will not have a finite radius. However, in some models the universe has finite spatial extent, in which case R is related to the maximum distance between two points in the universe.

It is helpful to consider the analogy of a spherical balloon which is expanding and which is uniformly covered on its surface with dots. The dots can be considered to correspond to 'galaxies' in a two-dimensional universe. As the balloon expands, all dots move away from each other and from any given dot all dots appear to move away with speeds which at any given time are proportional to the distance (along the surface). Let the

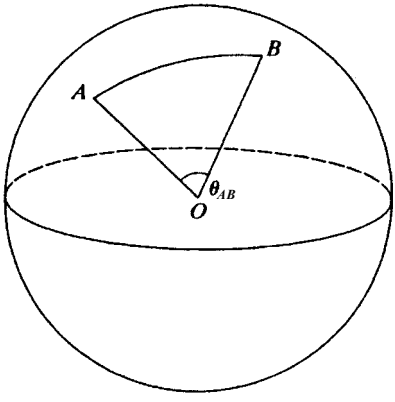


Fig. 1.1. Diagram to illustrate Equation (1.1).

radius of the balloon at time t be denoted by $R'(t)$. Consider two dots which subtend an angle θ_{AB} at the centre, the dots being denoted by A and B (Fig. 1.1). The distance d_{AB} between the dots on a great circle is given by

$$d_{AB} = \theta_{AB} R'(t). \tag{1.1}$$

The speed v_{AB} with which A and B are moving relative to each other is given by

$$v_{AB} = \dot{d}_{AB} = \theta_{AB} \dot{R}' = d_{AB}(\dot{R}'/R'), \quad \dot{R}' \equiv \frac{\mathrm{d}R'}{\mathrm{d}t}, \text{ etc.} \tag{1.2}$$

Thus the relative speed of A and B around a great circle is proportional to the distance around the great circle, the factor of proportionality being $\dot{R}'/R'$, which is the same for any pair of dots. The distance around a great circle between any pair of dots has the same form, for example, $\theta_{CD}R'$, where θ_{CD} is the angle subtended at the centre by dots C and D. Because the expansion of the balloon is uniform, the angles θ_{AB}, θ_{CD}, etc., remain the same for all t. We thus have a close analogy between the model of an expanding universe and the expansion of a uniformly dotted spherical balloon. In the case of galaxies Hubble's law is approximate but for dots on a balloon the corresponding relation is strictly true. From (1.1) it follows that if the distance between A and B changes by a certain factor in any period of time, the distance between *any* pair of dots changes by the same factor in that period of time.

From the rate at which galaxies are receding from each other, it can be deduced that *all* galaxies must have been very close to each other *at the same time* in the past. Considering again the analogy of the balloon, it is

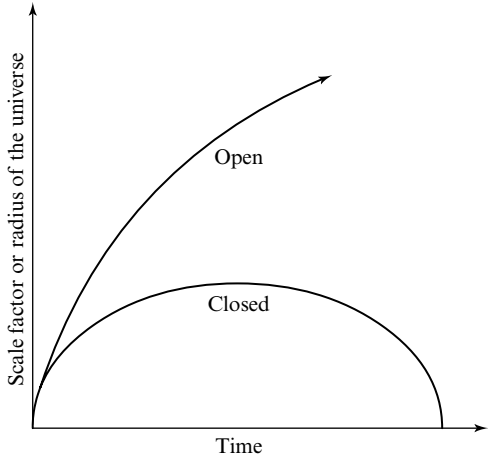


Fig. 1.2. Evolution of the scale factor or radius with time in the open and closed models of the universe.

like saying that the balloon must have started with zero radius and at this initial time all dots must have been on top of each other. For the universe it is believed that at this initial moment (some time between 10 and 20 billion years ago) there was a universal explosion, at every point of the universe, in which matter was thrown asunder violently. This was the 'big bang'. The explosion could have been at every point of an infinite or a finite universe. In the latter case the universe would have started from zero volume. An infinite universe remains infinite in spatial extent all the time down to the initial moment; as in the case of the finite universe, the matter becomes more and more dense and hot as one traces the history of the universe to the initial moment, which is a 'space-time singularity' about which we will learn more later. The universe is expanding now because of the initial explosion. There is not necessarily any force propelling the galaxies apart, but their motion can be explained as a remnant of the initial impetus. The recession is slowing down because of the gravitational attraction of different parts of the universe to each other, at least in the simpler models. This is not necessarily true in models with a cosmological constant, as we shall see later.

The expansion of the universe may continue forever, as in the 'open' models, or the expansion may halt at some future time and contraction set in, as in the 'closed' models, in which case the universe will collapse at a finite time later into a space-time singularity with infinite or near infinite density. These possibilities are illustrated in Fig. 1.2. In the Friedmann models the open universes have infinite spatial extent whereas the closed

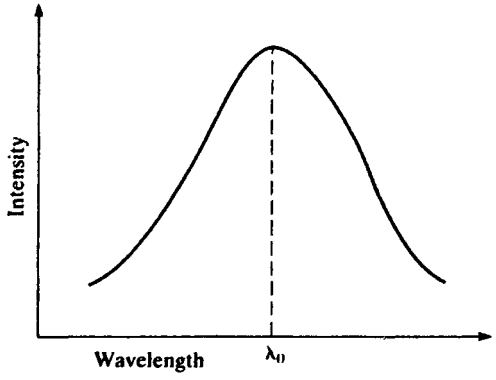


Fig. 1.3. Graph of intensity versus wavelength for black-body radiation. For the cosmic background radiation λ_0 is just under 0.1 cm.

models are finite. This is not necessarily the case for the Lemaître models. Both the Friedmann and Lemaître models will be discussed in detail in later chapters.

There is an important piece of evidence apart from the recession of the galaxies that the contents of the universe in the past must have been in a highly compressed form. This is the 'cosmic background radiation', which was discovered by Penzias and Wilson in 1965 and confirmed by many observations later. The existence of this radiation can be explained as follows. As we trace the history of the universe backwards to higher densities, at some stage galaxies could not have had a separate existence, but must have been merged together to form one great continuous mass. Due to the compression the temperature of the matter must have been very high. There is reason to believe, as we shall see, that there must also have been present a great deal of electromagnetic radiation, which at some stage was in equilibrium with the matter. The spectrum of the radiation would thus correspond to a black body of high temperature. There should be a remnant of this radiation, still with black-body spectrum, but corresponding to a much lower temperature. The cosmic background radiation discovered by Penzias, Wilson and others indeed does have a black-body spectrum (Fig. 1.3) with a temperature of about 2.7 K.

Hubble's law implies arbitrarily large velocities of the galaxies as the distance increases indefinitely. There is thus an apparent contradiction with special relativity which can be resolved as follows. The red-shift z is defined as $z = (\lambda_r - \lambda_i)/\lambda_i$, where λ_i is the original wavelength of the radiation given off by the galaxy and λ_r is the wavelength of this radiation when received

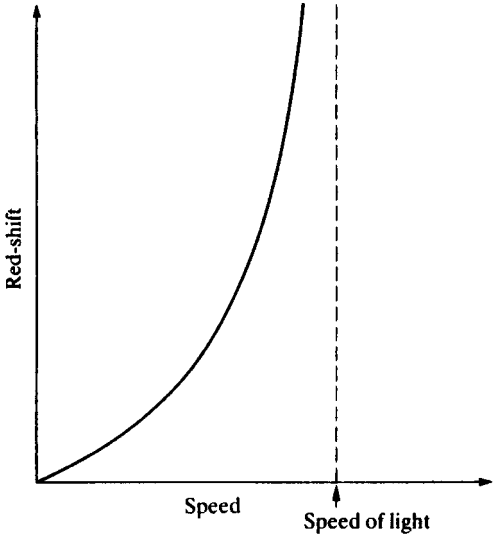


Fig. 1.4. This graph shows the relation between the red-shift (z) and the speed of recession. As z tends to infinity, the speed of recession tends to the speed of light.

by us. As the velocity of the galaxy approaches that of light, z tends towards infinity (Fig. 1.4), so it is not possible to *observe* higher velocities than that of light. The distance at which the red-shift of a galaxy becomes infinite is called the *horizon*. Galaxies beyond the horizon are indicated by Hubble's law to have higher velocities than light, but this does not violate special relativity because the presence of gravitation radically alters the nature of space and time according to general relativity. It is not as if a material particle is going past an observer at a velocity greater than that of light, but it is space which is in some sense expanding faster than the speed of light. This will become clear when we derive the expressions for the velocity, red-shift, etc., analytically later.

As mentioned earlier, in the open model the universe will expand forever whereas in the closed model there will be contraction and collapse in the future. It is not known at present whether the universe is open or closed. There are several interconnecting ways by which this could be determined. One way is to measure the present average density of the universe and compare it with a certain critical density. If the density is above the critical density, the attractive force of different parts of the universe towards each other will be enough to halt the recession eventually and to pull the galaxies together. If the density is below the critical density, the attractive force is

insufficient and the expansion will continue forever. The critical density at any time (this will be derived in detail later) is given by

$$\varepsilon_c = 3H^2/8\pi G, \quad H = \dot{R}/R. \tag{1.3}$$

Here G is Newton's gravitational constant and R is the scale factor which is a function of time; it corresponds to $R'(t)$ of (1.1) and represents the 'size' of the universe in a sense which will become clear later. If t_0 denotes the present time, then the present value of H, denoted by H_0, is called Hubble's constant. That is, $H_0 = H(t_0)$. For galaxies which are not too near nor too far, the velocity v is related to the distance d by Hubble's constant:

$$v = H_0 d. \tag{1.4}$$

(Compare (1.2), (1.3) and (1.4).) The present value of the critical density is thus $3H_0^2/8\pi G$, and is dependent on the value of Hubble's constant. There are some uncertainties in the value of the latter, the likely value being between 50 km s^{-1} and 100 km s^{-1} per million parsecs. That is, a galaxy which is 100 million parsecs distant has a velocity away from us of 5000–10 000 km s^{-1}. For a value of Hubble's constant given by 50 km s^{-1} per million parsecs, the critical density equals about 5×10^{-30} g cm^{-3}, or about three hydrogen atoms per thousand litres of space.

There are several other related ways of determining if the universe will expand forever. One of these is to measure the rate at which the expansion of the universe is slowing down. This is measured by the deceleration parameter, about which there are also uncertainties. Theoretically in the simpler models, in suitable units, the deceleration parameter is half the ratio of the actual density to the critical density. This ratio is usually denoted by Ω. Thus if $\Omega < 1$, the density is subcritical and the universe will expand forever, the opposite being the case if $\Omega > 1$. The present observed value of Ω is somewhere between 0.1 and 2 (the lower limit could be less). In the simpler models the deceleration parameter, usually denoted by q_0, is thus $\frac{1}{2}\Omega$, so that the universe expands forever in these models if $q_0 < \frac{1}{2}$, the opposite being the case if $q_0 > \frac{1}{2}$.

Another way to find out if the universe will expand forever is to determine the precise age of the universe and compare it with the 'Hubble time'. This is the time elapsed since the big bang until now if the rate of expansion had been the same as at present. In Fig. 1.5 if ON denotes the present time (t_0), then clearly PN is $R(t_0)$. If the tangent at P to the curve $R(t)$ meets the t-axis at T at an angle α, then

$$\tan\alpha = PN/NT = \dot{R}(t_0), \tag{1.5}$$

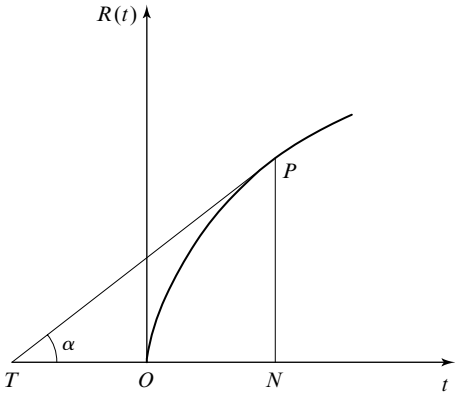


Fig. 1.5. Diagram to define Hubble time.

so that

$$NT = PN/\dot{R}(t_0) = R(t_0)/\dot{R}(t_0) = H_0^{-1}. \tag{1.6}$$

Thus NT, which is, in fact, Hubble's time, is the reciprocal of Hubble's constant in the units considered here. For the value of 50 km s^{-1} per million parsecs of Hubble's constant, the Hubble time is about 20 billion years. Again in the simpler models, if the universe is older than two-thirds of the Hubble time it will expand forever, the opposite being the case if its age is less than two-thirds of the Hubble time.

Whether the universe will expand forever is one of the most important unresolved problems in cosmology, both theoretically and observationally, but all the above methods of ascertaining this contain many uncertainties.

In this book we shall use the term 'open' to mean a model which expands forever, and 'closed' for the opposite. Sometimes the expression 'closed' is used to mean a universe with a finite volume, but, as mentioned earlier, it is only in the Friedmann models that a universe has infinite volume if it expands forever, etc.

The standard big-bang model of the universe has had three major successes. Firstly, it predicts that something like Hubble's law of expansion must hold for the universe. Secondly, it predicts the existence of the microwave background radiation. Thirdly, it predicts successfully the formation of light atomic nuclei from protons and neutrons a few minutes after the big bang. This prediction gives the correct abundance ratio for He^3, D, He^4 and Li^7. (We shall discuss this in detail later.) Heavier elements are thought

to have been formed much later in the interior of stars. (See Hoyle, Burbidge and Narlikar (2000) for an alternative point of view.)

Certain problems and puzzles remain in the standard model. One of these is that the universe displays a remarkable degree of large-scale homogeneity. This is most evident in the microwave background radiation which is known to be uniform in temperature to about one part in 1000. (There is, however, a systematic variation of about one part in 3000 attributed to the motion of the Earth in the Galaxy and the motion of the Galaxy in the local group of galaxies, and also a smaller variation in all directions, presumably due to the 'graininess' that existed in the matter at the time the radiation 'decoupled'.) The uniformity that exists is a puzzle because, soon after the big bang, regions which were well separated could not have communicated with each other or known of each other's existence. Roughly speaking, at a time t after the big bang, light could have travelled only a distance ct since the big bang, so regions separated by a distance greater than ct at time t could not have influenced each other. The fact that microwave background radiation received from all directions is uniform implies that there is uniformity in regions whose separation must have been many times the distance ct (the *horizon distance*) a second or so after the big bang. How did these different regions manage to have the same density, etc.? Of course there is no problem if one simply *assumes* that the uniformity persists up to time $t = 0$, but this requires a very special set of initial conditions. This is known as the *horizon problem*.

Another problem is concerned with the fact that a certain amount of inhomogeneity must have existed in the primordial matter to account for the clumping of matter into galaxies and clusters of galaxies, etc., that we observe today. Any small inhomogeneity in the primordial matter rapidly grows into a large one with gravitational self-interaction. Thus one has to assume a considerable smoothness in the primordial matter to account for the inhomogeneity in the scale of galaxies at the present time. The problem becomes acute if one extrapolates to 10^{-45} s after the big bang, when one has to assume an unusual situation of almost perfect smoothness but not quite absolute smoothness in the initial state of matter. This is known as the *smoothness problem*.

A third problem of the standard big-bang model has to do with the present observed density of matter, which we have denoted by the parameter Ω. If Ω were initially equal to unity (this corresponds to a flat universe) it would stay equal to unity forever. On the other hand, if Ω were initially different from unity, its depature from unity would increase with time. The present value of Ω lies somewhere between 0.1 and 2. For this to be the

case the value of Ω would have had to be equal to 1 to one part in 10^{15} a second or so after the big bang, which seems an unlikely situation. This is called the *flatness problem*.

To deal with these problems Alan Guth (1981) proposed a model of the universe, known as the inflationary model, which does not differ from the standard model after a fraction of a second or so, but from about 10^{-45} to 10^{-30} seconds it has a period of extraordinary expansion, or inflation, during which time typical distances (the scale factor) increase by a factor of about 10^{50} more than the increase that would obtain in the standard model. Although the inflationary models (there have been variations of the one put forward by Guth originally) solve some of the problems of the standard models, they throw up problems of their own, which have not all been dealt with in a satisfactory manner. These models will be considered in detail in this book.

The consideration of the universe in the first second or so calls for a great deal of information from the theory of elementary particles, particularly in the inflationary models. This period is referred to as 'the very early universe' and it also provides a testing ground for various theories of elementary particles. These questions will be considered in some detail in a later chapter.

As one extrapolates in time to the very early universe and towards the big bang at $t = 0$, densities become higher and higher and the curvature of space-time becomes correspondingly higher, and at some stage general relativity becomes untenable and one has to resort to the quantum theory of gravitation. However, a satisfactory quantum theory for gravity does not yet exist. Some progress has been made in what is called 'quantum cosmology', in which quantum considerations throw some light on problems to do with initial conditions of the universe. We shall attempt to provide an introduction to this subject in this book.

If the universe is open, that is, if it expands for ever, one has essentially infinite time in the future for the universe to evolve. What will be the nature of this evolution and what will be the final state of the universe? These questions and related ones will be considered in Chapter 11.

2

Introduction to general relativity

2.1 Summary of general relativity

The Robertson–Walker metric or line-element is fundamental in the standard models of cosmology. The mathematical framework in which the Robertson–Walker metric occurs is that of general relativity. The reader is assumed to be familiar with general relativity but we shall give an introduction here as a reminder of the main results and for the sake of completeness. We shall then go on to derive the Robertson–Walker metric in the next chapter. We begin with a brief summary.

General relativity is formulated in a four-dimensional Riemannian space in which points are labelled by a general coordinate system (x^0, x^1, x^2, x^3), often written as x^μ ($\mu = 0, 1, 2, 3$). (Greek indices take values of 0, 1, 2, 3 and repeated Greek indices are to be summed over these values.) Several coordinate patches may be necessary to cover the whole of space-time. The space has three spatial and one time-like dimension.

Under a coordinate transformation from x^μ to x'^μ (in which x'^μ is, in general, a function of x^0, x^1, x^2, x^3) a contravariant vector field A^μ and a covariant vector field B_μ transform as follows:

$$A'^\mu = \frac{\partial x'^\mu}{\partial x^\nu} A^\nu, \quad B'_\mu = \frac{\partial x^\nu}{\partial x'^\mu} B_\nu, \tag{2.1}$$

and a mixed tensor such as $A^\mu_{\nu\lambda}$ transforms as follows:

$$A'^\mu_{\nu\lambda} = \frac{\partial x'^\mu}{\partial x^\rho} \frac{\partial x^\sigma}{\partial x'^\nu} \frac{\partial x^\tau}{\partial x'^\lambda} A^\rho_{\sigma\tau}, \tag{2.2}$$

etc. All the information about the gravitational field is contained in the second rank covariant tensor $g_{\mu\nu}$ (the number of indices gives the rank of

the tensor) called the metric tensor, or simply the metric, which determines the square of the space-time intervals ds^2 between infinitesimally separated events or points x^μ and $x^\mu + dx^\mu$ as follows ($g_{\mu\nu} = g_{\nu\mu}$):

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu. \tag{2.3}$$

The contravariant tensor corresponding to $g_{\mu\nu}$ is denoted by $g^{\mu\nu}$ and is defined by

$$g_{\mu\nu} g^{\nu\lambda} = \delta^\lambda_\mu, \tag{2.4}$$

where δ^λ_μ is the Kronecker delta, which equals unity if $\lambda = \mu$ (no summation) and zero otherwise. Indices can be raised or lowered by using the metric tensor as follows:

$$A^\mu = g^{\mu\nu} A_\nu, \quad A_\mu = g_{\mu\nu} A^\nu. \tag{2.5}$$

The generalization of ordinary (partial) differentiation to Riemannian space is given by covariant differentiation denoted by a semi-colon and defined for a contravariant and a covariant vector as follows:

$$A^\mu{}_{;\nu} = \frac{\partial A^\mu}{\partial x^\nu} + \Gamma^\mu_{\nu\lambda} A^\lambda, \tag{2.6a}$$

$$A_{\mu;\nu} = \frac{\partial A_\mu}{\partial x^\nu} - \Gamma^\lambda_{\mu\nu} A_\lambda. \tag{2.6b}$$

Here the $\Gamma^\mu_{\nu\lambda}$ are called Christoffel symbols; they have the property $\Gamma^\mu_{\nu\lambda} = \Gamma^\mu_{\lambda\nu}$ and are given in terms of the metric tensor as follows:

$$\Gamma^\mu_{\nu\lambda} = \tfrac{1}{2} g^{\mu\sigma} (g_{\sigma\nu,\lambda} + g_{\sigma\lambda,\nu} - g_{\nu\lambda,\sigma}), \tag{2.7}$$

where a comma denotes partial differentiation with respect to the corresponding variable: $g_{\sigma\nu,\lambda} \equiv \partial g_{\sigma\nu} / \partial x^\lambda$. For covariant differentiation of tensors of higher rank, there is a term corresponding to each contravariant index analogous to the second term in (2.6a) and a term corresponding to each covariant index analogous to the second term in (2.6b) (with a negative sign). For example, the covariant derivative of the mixed tensor considered in (2.2) can be written as follows:

$$A^\mu{}_{\nu\lambda;\sigma} = \frac{\partial A^\mu_{\nu\lambda}}{\partial x^\sigma} + \Gamma^\mu_{\sigma\rho} A^\rho_{\nu\lambda} - \Gamma^\rho_{\nu\sigma} A^\mu_{\rho\lambda} - \Gamma^\rho_{\lambda\sigma} A^\mu_{\nu\rho}. \tag{2.6c}$$

Equation (2.7) has the consequence that the covariant derivative of the metric tensor vanishes:

$$g_{\mu\nu;\lambda} = 0, \quad g^{\mu\nu}{}_{;\lambda} = 0. \tag{2.8}$$

This has, in turn, the consequence that indices can be raised and lowered inside the sign for covariant differentiation, as follows:

$$g_{\sigma\mu}A^{\mu}{}_{;\nu} = A_{\sigma;\nu}, \quad g^{\sigma\mu}A_{\mu;\nu} = A^{\sigma}{}_{;\nu}. \tag{2.9}$$

Under a coordinate transformation from x^{μ} to x'^{μ} the $\Gamma^{\mu}_{\nu\lambda}$ transform follows:

$$\Gamma'^{\mu}_{\nu\lambda} = \frac{\partial x'^{\mu}}{\partial x^{\rho}} \frac{\partial x^{\sigma}}{\partial x'^{\nu}} \frac{\partial x^{\tau}}{\partial x'^{\lambda}} \Gamma^{\rho}_{\sigma\tau} + \frac{\partial^2 x^{\sigma}}{\partial x'^{\nu}\, \partial x'^{\lambda}} \frac{\partial x'^{\mu}}{\partial x^{\sigma}}, \tag{2.10}$$

so that the $\Gamma^{\mu}_{\nu\lambda}$ do not form components of a tensor since the transformation law (2.10) is different from that of a tensor (see (2.2)). At any specific point a coordinate system can always be chosen so that the $\Gamma^{\mu}_{\nu\lambda}$ vanish at the point. From (2.7) it follows that the first derivatives of the metric tensor also vanish at this point. This is one form of the equivalence principle, according to which the gravitational field can be 'transformed away' at any point by choosing a suitable frame of reference. At this point one can carry out a further linear transformation of the coordinates to reduce the metric to that of flat (Minkowski) space:

$$\mathrm{d}s^2 = (\mathrm{d}x^0)^2 - (\mathrm{d}x^1)^2 - (\mathrm{d}x^2)^2 - (\mathrm{d}x^3)^2, \tag{2.11}$$

where $x^0 = ct$, t being the time and (x^1, x^2, x^3) being Cartesian coordinates.

For any covariant vector A_{μ} it can be shown that

$$A_{\mu;\nu;\lambda} - A_{\mu;\lambda;\nu} = A_{\sigma} R^{\sigma}_{\mu\nu\lambda}, \tag{2.12}$$

where $R^{\sigma}_{\mu\nu\lambda}$ is the Riemann tensor defined by

$$R^{\sigma}_{\mu\nu\lambda} = \Gamma^{\sigma}_{\mu\lambda,\nu} - \Gamma^{\sigma}_{\mu\nu,\lambda} + \Gamma^{\sigma}_{\alpha\nu}\Gamma^{\alpha}_{\mu\lambda} - \Gamma^{\sigma}_{\alpha\lambda}\Gamma^{\alpha}_{\mu\nu}. \tag{2.13}$$

The Riemann tensor has the following symmetry properties:

$$R_{\sigma\mu\nu\lambda} = -R_{\mu\sigma\nu\lambda} = -R_{\sigma\mu\lambda\nu}, \tag{2.14a}$$

$$R_{\sigma\mu\nu\lambda} = R_{\nu\lambda\sigma\mu}, \tag{2.14b}$$

$$R_{\sigma\mu\nu\lambda} + R_{\sigma\lambda\mu\nu} + R_{\sigma\nu\lambda\mu} = 0, \tag{2.14c}$$

and satisfies the Bianchi identity:

$$R^{\sigma}_{\mu\nu\lambda;\rho} + R^{\sigma}_{\mu\rho\nu;\lambda} + R^{\sigma}_{\mu\lambda\rho;\nu} = 0. \tag{2.15}$$

The Ricci tensor $R_{\mu\nu}$ is defined by

$$R_{\mu\nu} = g^{\lambda\sigma} R_{\lambda\mu\sigma\nu} = R^{\sigma}_{\mu\sigma\nu}. \tag{2.16}$$

From (2.13) and (2.16) it follows that $R_{\mu\nu}$ is given as follows:

$$R_{\mu\nu} = \Gamma^{\lambda}_{\mu\nu,\lambda} - \Gamma^{\lambda}_{\mu\lambda,\nu} + \Gamma^{\lambda}_{\mu\nu}\Gamma^{\sigma}_{\lambda\sigma} - \Gamma^{\sigma}_{\mu\lambda}\Gamma^{\lambda}_{\nu\sigma}. \tag{2.17}$$

Let the determinant of $g_{\mu\nu}$ considered as a matrix be denoted by g. Then another expression for $R_{\mu\nu}$ is given by the following:

$$R_{\mu\nu} = \frac{1}{(-g)^{1/2}}[\Gamma^{\lambda}_{\mu\nu}(-g)^{1/2}]_{,\lambda} - [\log(-g)^{1/2}]_{,\mu\nu} - \Gamma^{\sigma}_{\mu\lambda}\Gamma^{\lambda}_{\nu\sigma}. \tag{2.18}$$

This follows from the fact that from (2.7) and the properties of matrices one can show that

$$\Gamma^{\lambda}_{\mu\lambda} = [\log(-g)^{1/2}]_{,\mu}. \tag{2.19}$$

From (2.18) it follows that $R_{\mu\nu} = R_{\nu\mu}$. There is no agreed convention for the signs of the Riemann and Ricci tensors – some authors define these with opposite signs to (2.13) and (2.17). The Ricci scalar R is defined by

$$R = g^{\mu\nu}R_{\mu\nu}. \tag{2.20}$$

By contracting the Bianchi identity (2.15) on the pair of indices $\mu\nu$ and $\sigma\rho$ (that is, multiplying it by $g^{\mu\nu}$ and $g^{\sigma\rho}$) one can deduce the identity

$$(R^{\mu\nu} - \tfrac{1}{2}g^{\mu\nu}R)_{;\nu} = 0. \tag{2.21}$$

The tensor $G^{\mu\nu} = R^{\mu\nu} - \frac{1}{2}g^{\mu\nu}R$ is sometimes called the Einstein tensor.

We are now in a position to write down the fundamental equations of general relativity. These are Einstein's equations given by:

$$R_{\mu\nu} - \tfrac{1}{2}g_{\mu\nu}R = (8\pi G/c^4)T_{\mu\nu}, \tag{2.22}$$

where $T_{\mu\nu}$ is the energy–momentum tensor of the source producing the gravitational field and G is Newton's gravitational constant. For a perfect fluid, $T_{\mu\nu}$ takes the following form:

$$T^{\mu\nu} = (\varepsilon + p)u^{\mu}u^{\nu} - pg^{\mu\nu}, \tag{2.23}$$

where ε is the mass-energy density, p is the pressure and u^{μ} is the four-velocity of matter given by

$$u^{\mu} = \frac{\mathrm{d}x^{\mu}}{\mathrm{d}s}, \tag{2.24}$$

where $x^{\mu}(s)$ describes the worldline of matter in terms of the proper time $\tau = c^{-1}s$ along the worldline. We will consider later some other forms of the energy–momentum tensor than (2.23). From (2.21) we

see that Einstein's equations (2.22) are compatible with the following equation

$$T^{\mu\nu}{}_{;\nu} = 0, \tag{2.25}$$

which is the equation for the conservation of mass-energy and momentum.

The equations of motion of a particle in a gravitational field are given by the geodesic equations as follows:

$$\frac{\mathrm{d}^2 x^\mu}{\mathrm{d}s^2} + \Gamma^\mu_{\lambda\nu}\frac{\mathrm{d}x^\lambda}{\mathrm{d}s}\frac{\mathrm{d}x^\nu}{\mathrm{d}s} = 0. \tag{2.26}$$

Geodesics can also be introduced through the concept of parallel transfer. Consider a curve $x^\mu(\lambda)$, where x^μ are suitably differentiable functions of the real parameter λ, varying over some interval of the real line. It is readily verified that $\mathrm{d}x^\mu/\mathrm{d}\lambda$ transforms as a contravariant vector. This is the tangent vector to the curve $x^\mu(\lambda)$. For an arbitrary vector field Y^μ its covariant derivative along the curve (defined along the curve) is $Y^\mu{}_{;\nu}(\mathrm{d}x^\nu/\mathrm{d}\lambda)$. The vector field Y^μ is said to be parallelly transported along the curve if

$$\begin{aligned} Y^\mu{}_{;\nu}\frac{\mathrm{d}x^\nu}{\mathrm{d}\lambda} &= Y^\mu{}_{,\nu}\frac{\mathrm{d}x^\nu}{\mathrm{d}\lambda} + \Gamma^\mu_{\nu\sigma}Y^\sigma\frac{\mathrm{d}x^\nu}{\mathrm{d}\lambda} \\ &= \frac{\mathrm{d}Y^\mu}{\mathrm{d}\lambda} + \Gamma^\mu_{\nu\sigma}Y^\sigma\frac{\mathrm{d}x^\nu}{\mathrm{d}\lambda} = 0. \end{aligned} \tag{2.27}$$

The curve is said to be a geodesic curve if the tangent vector is transported parallelly along the curve, that is, putting ($Y^\mu = \mathrm{d}x^\mu/\mathrm{d}\lambda$ in (2.27)) if

$$\frac{\mathrm{d}^2 x^\mu}{\mathrm{d}\lambda^2} + \Gamma^\mu_{\nu\sigma}\frac{\mathrm{d}x^\nu}{\mathrm{d}\lambda}\frac{\mathrm{d}x^\sigma}{\mathrm{d}\lambda} = 0. \tag{2.28}$$

The curve, or a portion of it, is time-like, light-like or space-like according as to whether $g_{\mu\nu}(\mathrm{d}x^\mu/\mathrm{d}\lambda)(\mathrm{d}x^\nu/\mathrm{d}\lambda) > 0, = 0$, or < 0. (As mentioned earlier, at any point $g_{\mu\nu}$ can be reduced to the diagonal form $(1,-1,-1,-1)$ by a suitable transformation.) The length of the time-like or space-like curve from $\lambda = \lambda_1$ to $\lambda = \lambda_2$ is given by:

$$L_{12} = \int_{\lambda_1}^{\lambda_2}\left(\left|g_{\mu\nu}\frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda}\frac{\mathrm{d}x^\nu}{\mathrm{d}\lambda}\right|\right)^{1/2}\mathrm{d}\lambda. \tag{2.29}$$

If the tangent vector $\mathrm{d}x^\mu/\mathrm{d}\lambda$ is time-like everywhere, the curve $x^\mu(\lambda)$ can be taken to be the worldline of a particle and λ the proper time $c^{-1}s$

along the worldline, and in this case (2.28) reduces to (2.26). The former equation has more general applicability, for example, when the curve $x^\mu(\lambda)$ is light-like or space-like, in which case λ cannot be taken as the proper time.

Two vector fields V^μ, W^μ are normal or orthogonal to each other if $g_{\mu\nu}V^\mu W^\nu = 0$. If V^μ is time-like and orthogonal to W^μ then the latter is necessarily space-like. A space-like three-surface is a surface defined by $f(x^0,x^1,x^2,x^3) = 0$ such that $g^{\mu\nu}f_{,\mu}f_{,\nu} > 0$ when $f = 0$. The unit normal vector to this surface is given by $n^\mu = (g^{\alpha\beta}f_{,\alpha}f_{,\beta})^{-1/2}\, g^{\mu\nu}f_{,\nu}$.

Given a vector field ζ^μ, one can define a set of curves filling all space such that the tangent vector to any curve of this set at any point coincides with the value of the vector field at that point. This is done by solving the set of first order differential equations.

$$\frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda} = \zeta^\mu(x(\lambda)), \tag{2.30}$$

where on the right hand side we have put x for all four components of the coordinates. This set of curves is referred to as the congruence of curves generated by the given vector field. In general there is a unique member of this congruence passing through any given point. A particular member of the congruence is sometimes referred to as an orbit. Consider now the vector field given by $(\zeta^0,\zeta^1,\zeta^2,\zeta^3) = (1,0,0,0)$. From (2.30) we see that the congruence of this vector field is the set of curves given by

$$(x^0 = \lambda, x^1 = \text{constant}, x^2 = \text{constant}, x^3 = \text{constant}). \tag{2.31}$$

This vector field is also referred to as the vector field $\partial/\partial x^0$. One similarly defines the vector fields $\partial/\partial x^1$, $\partial/\partial x^2$, $\partial/\partial x^3$. That is, corresponding to the coordinate system x^μ we have the four contravariant vector fields $\partial/\partial x^\mu$. A general vector field X^μ can be written without components in terms of $\partial/\partial x^\mu$ as follows:

$$\mathrm{X} = X^\mu \frac{\partial}{\partial x^\mu}. \tag{2.32}$$

This is related to the fact that contravariant vectors at any point can be regarded as operators acting on differentiable functions $f(x^0,x^1,x^2,x^3)$; when the vector acts on the function, the result is the derivative of the function in the direction of the vector field, as follows:

$$\mathrm{X}(f) = X^\mu \frac{\partial f}{\partial x^\mu}. \tag{2.33}$$

As is well known, differential geometry and, correspondingly, general relativity can be developed independently of coordinates and components. We shall not be concerned with this approach except incidentally (see, for example, Hawking and Ellis, 1973).

We will now consider some special topics in general relativity which may not all be used directly in the following chapters, but which may be useful in some contexts in cosmological studies.

2.2 Some special topics in general relativity

2.2.1 Killing vectors

Einstein's exterior equations $R_{\mu\nu}=0$ (obtained from (2.22) by setting $T_{\mu\nu}=0$) are a set of coupled non-linear partial differential equations for the ten unknown functions $g_{\mu\nu}$. The interior equations (2.22) may involve other unknown functions such as the mass-energy density and the pressure. Because of the freedom to carry out general coordinate transformations one can in general impose four conditions on the ten functions $g_{\mu\nu}$. Later we will show explicitly how this is done in a case involving symmetries. In most situations of physical interest one has space-time symmetries which reduce further the number of unknown functions. To determine the simplest form of the metric (that is, the form of $g_{\mu\nu}$) when one has a given space-time symmetry is a non-trivial problem. For example, in Newtonian theory spherical symmetry is usually defined by a centre and the property that all points at any given distance from the centre are equivalent. This definition cannot be taken over directly to general relativity. In the latter, 'distance' is defined by the metric to begin with and, for example, the 'centre' may not be accessible to physical measurement, as is indeed the case in the Schwarzschild geometry (see Section 7.4). One therefore has to find some coordinate independent and covariant manner of defining space-time symmetries such as axial symmetry and stationarity. This is done with the help of Killing vectors, which we will now consider. In some cases there is a less rigorous but simpler way of deriving the metric which we will also consider.

In the following we will sometimes write x, y, x' for x^μ, y^μ, x'^μ respectively. A metric $g_{\mu\nu}(x)$ is form-invariant under a transformation from x^μ to x'^μ if $g'_{\mu\nu}(x')$ is the same function of x'^μ as $g_{\mu\nu}(x)$ is of x^μ. For example, the Minkowski metric is form-invariant under a Lorentz transformation. Thus

$$g'_{\mu\nu}(y)=g_{\mu\nu}(y), \quad \text{all } y. \tag{2.34}$$

Therefore

$$g_{\mu\nu}(x) = \frac{\partial x'^{\rho}}{\partial x^{\mu}} \frac{\partial x'^{\sigma}}{\partial x^{\nu}} g'_{\rho\sigma}(x') = \frac{\partial x'^{\rho}}{\partial x^{\mu}} \frac{\partial x'^{\sigma}}{\partial x^{\nu}} g_{\rho\sigma}(x'). \tag{2.35}$$

The transformation from x^{μ} to x'^{μ} in this case is called an isometry of $g_{\mu\nu}$. Consider an infinitesimal isometry transformation from x^{μ} to x'^{μ} defined by

$$x'^{\mu} = x^{\mu} + \alpha\xi^{\mu}(x), \tag{2.36}$$

with α constant and $|\alpha| \ll 1$. Substituting in (2.35) and neglecting terms involving α^2 we arrive at the following equation (see e.g. Weinberg (1972)):

$$g_{\mu\sigma}\frac{\partial \xi^{\mu}}{\partial x^{\rho}} + g_{\rho\mu}\frac{\partial \xi^{\mu}}{\partial x^{\sigma}} + \frac{\partial g_{\rho\sigma}}{\partial x^{\mu}}\xi^{\mu} = 0. \tag{2.37}$$

With the use of (2.6b) and (2.7) the equation (2.37) can be written as follows:

$$\xi_{\sigma;\rho} + \xi_{\rho;\sigma} = 0. \tag{2.38}$$

Equation (2.38) is Killing's equation and a vector field ξ^{μ} satisfying it is called a Killing vector of the metric $g_{\mu\nu}$. Thus if there exists a solution of (2.38) for a given $g_{\mu\nu}$, then the corresponding ξ^{μ} represents an infinitesimal isometry of the metric $g_{\mu\nu}$ and implies that the metric has a certain symmetry. Since (2.38) is covariantly expressed, that is, it is a tensor equation, if the metric has an isometry in a given coordinate system, in any transformed coordinate system the transformed metric will also have a corresponding isometry. This is important because often a metric can look quite different in different coordinate systems.

To give an example of a Killing vector, we consider a situation in which the metric is independent of one of the four coordinates. To fix ideas, we choose this coordinate to be x^0, which we take to be time-like, that is, the lines ($x^0 = \lambda$, $x^1 =$ constant, $x^2 =$ constant, $x^3 =$ constant) for varying λ are time-like lines. In general, $g_{\mu\nu}$ being independent of x^0 means that the gravitational field is stationary, that is, it is produced by sources whose state of motion does not change with time. In this case we have

$$\frac{\partial g_{\mu\nu}}{\partial x^0} \equiv g_{\mu\nu,0} = 0. \tag{2.39}$$

Consider now the vector field ξ^{μ} given by

$$(\xi^0, \xi^1, \xi^2, \xi^3) = (1,0,0,0), \tag{2.40}$$

with $\xi_\mu = g_{\mu\nu}\xi^\nu = g_{\mu 0}$. We have

$$\begin{aligned}\xi_{\mu;\nu} + \xi_{\nu,\mu} &= \xi_{\mu,\nu} + \xi_{\nu,\mu} - g^{\lambda\sigma}(g_{\sigma\mu,\nu} + g_{\sigma\nu,\mu} - g_{\mu\nu,\sigma})\xi_\lambda \\ &= g_{\mu 0,\nu} + g_{\nu 0,\mu} - \xi^\sigma(g_{\sigma\mu,\nu} + g_{\sigma\nu,\mu} - g_{\mu\nu,\sigma}) \\ &= g_{\mu\nu,0} = 0, \end{aligned} \tag{2.41}$$

using (2.39) and (2.40). Thus if (2.39) is satisfied, the vector (2.40) gives a solution to Killing's equation. In other words, if the metric admits the Killing vector (2.40), then (2.39) is satisfied and the metric is stationary. A similar result can be established for any of the other three coordinates.

We now derive a property of Killing vectors which we will use later. Let $\xi^{(1)\mu}$ and $\xi^{(2)\mu}$ be two linearly independent solutions of Killing's equation (2.38). We define the commutator of these two Killing vectors as the vector ζ^μ given by

$$\zeta^\mu = \xi^{(1)\mu}{}_{;\lambda}\xi^{(2)\lambda} - \xi^{(2)\mu}{}_{;\lambda}\xi^{(1)\lambda}. \tag{2.42}$$

In coordinate independent notation the commutator of $\xi^{(1)}$ and $\xi^{(2)}$ is written as $[\xi^{(1)}, \xi^{(2)}]$. In fact, because of the symmetry of the Christoffel symbols the covariant derivatives in (2.42) can be replaced by ordinary derivatives. We will now show that ζ^μ is also a Killing vector, that is,

$$\zeta_{\mu;\nu} + \zeta_{\nu;\mu} = 0. \tag{2.43}$$

Now

$$\begin{aligned}\zeta_{\mu;\nu} + \zeta_{\nu;\mu} = {} & \xi^{(1)}_{\mu;\lambda;\nu}\xi^{(2)\lambda} + \xi^{(1)}_{\mu;\lambda}\xi^{(2)\lambda}{}_{;\nu} - \xi^{(2)}_{\mu;\lambda;\nu}\xi^{(1)\lambda} \\ & - \xi^{(2)}_{\mu;\lambda}\xi^{(1)\lambda}{}_{;\nu} + \xi^{(1)\lambda}_{\nu;\lambda;\mu}\xi^{(2)\lambda} + \xi^{(1)}_{\nu;\lambda}\xi^{(2)\lambda}{}_{;\mu} \\ & - \xi^{(2)}_{\nu;\lambda;\mu}\xi^{(1)\lambda} - \xi^{(2)}_{\nu;\lambda}\xi^{(1)\lambda}{}_{;\mu}. \end{aligned} \tag{2.44}$$

From the fact that $\xi^{(1)\mu}$, $\xi^{(2)\mu}$ are Killing vectors, we have

$$\xi^{(i)}_{\mu;\nu;\lambda} + \xi^{(i)}_{\nu;\mu;\lambda} = 0, \quad i = 1, 2, \tag{2.45}$$

by taking the covariant derivative of Killing's equation. Also, from (2.12) we find that

$$\xi^{(i)}_{\mu;\nu;\lambda} = \xi^{(i)}_{\mu;\lambda;\nu} + \xi^{(i)\sigma}R_{\sigma\mu\nu\lambda}, \quad i = 1, 2. \tag{2.46}$$

With the use of (2.45) and (2.46) one can show that

$$\xi^{(1)}_{\mu;\lambda;\nu}\xi^{(2)\lambda} + \xi^{(1)}_{\nu;\lambda;\mu}\xi^{(2)\lambda} = \xi^{(1)\sigma}\xi^{(2)\lambda}(R_{\sigma\mu\lambda\nu} + R_{\sigma\nu\lambda\mu}), \tag{2.47a}$$

$$\xi^{(2)}_{\mu;\lambda;\nu}\xi^{(1)\lambda} + \xi^{(2)}_{\nu;\lambda;\mu}\xi^{(1)\lambda} = \xi^{(2)\sigma}\xi^{(1)\lambda}(R_{\sigma\mu\lambda\nu} + R_{\sigma\nu\lambda\mu}). \tag{2.47b}$$

Subtracting (2.47b) from (2.47a) we get

$$(\xi^{(1)}_{\mu;\lambda;\nu}+\xi^{(1)}_{\nu;\lambda;\mu})\xi^{(2)\lambda}-(\xi^{(2)}_{\mu;\lambda;\nu}+\xi^{(2)}_{\nu;\lambda;\mu})\xi^{(1)\lambda}$$
$$=\xi^{(1)\sigma}\xi^{(2)\lambda}(R_{\sigma\mu\lambda\nu}+R_{\sigma\nu\lambda\mu}-R_{\lambda\mu\sigma\nu}-R_{\lambda\nu\sigma\mu})=0, \tag{2.48}$$

where the last step follows from the symmetry properties of the Riemann tensor. Thus the terms on the right hand side of (2.44) involving double covariant derivatives vanish. The other terms can be shown to cancel by using Killing's equation. For example,

$$\begin{aligned}\xi^{(1)}_{\mu;\lambda}\xi^{(2)\lambda}{}_{;\nu}&=-\xi^{(1)}_{\lambda;\mu}\xi^{(2)\lambda}{}_{;\nu}\\&=-\xi^{(1)\lambda}{}_{;\mu}\xi^{(2)}_{\lambda;\nu}\\&=+\xi^{(1)\lambda}{}_{;\mu}\xi^{(2)}_{\nu;\lambda}\end{aligned} \tag{2.49}$$

which cancels the last term in (2.44), and so on. Thus ζ^μ satisfies (2.43) and so is a Killing vector. Suppose we have only n linearly independent Killing vectors $\xi^{(i)\mu}$, $i=1, 2, \ldots, n$ and no more. Then the commutator of any two of these is a Killing vector and so must be a linear combination of some or all of the n Killing vectors with constant coefficients since there are no other solutions of Killing's equation. Thus we have the result

$$\xi^{(i)\mu}{}_{;\nu}\xi^{(j)\nu}-\xi^{(j)\mu}{}_{;\nu}\xi^{(i)\nu}=\sum_{k=1}^{n}a^{ij}_k\xi^{(k)\mu},\quad i,j=1,\ldots,n. \tag{2.50}$$

In coordinate independent notation, we can write

$$[\xi^{(i)},\xi^{(j)}]=\sum_{k=1}^{n}a^{ij}_k\xi^{(k)},\quad i,j=1,\ldots,n. \tag{2.51}$$

In these two equations a^{ij}_k are constants.

2.2.2 Tensor densities

Tensor densities are needed in some contexts, such as volume and surface integrals. The latter are used in formulating an action principle from which field equations can be derived in a convenient manner. We shall use this principle to obtain the field equations with a scalar (Higgs) field in connection with inflationary cosmologies.

Consider a transformation from coordinates x^μ to x'^μ. An element of four-dimensional volume transforms as follows:

$$\mathrm{d}x'^0\mathrm{d}x'^1\mathrm{d}x'^2\mathrm{d}x'^3=J\mathrm{d}x^0\mathrm{d}x^1\mathrm{d}x^2\mathrm{d}x^3, \tag{2.52}$$

where J is the Jacobian of the transformation given by

$$J=\frac{\partial(x'^0,x'^1,x'^2,x'^3)}{\partial(x^0,x^1,x^2,x^3)}\equiv\begin{vmatrix}\frac{\partial x'^0}{\partial x^0} & \frac{\partial x'^0}{\partial x^1} & \frac{\partial x'^0}{\partial x^2} & \frac{\partial x'^0}{\partial x^3}\\ \frac{\partial x'^1}{\partial x^0} & \frac{\partial x'^1}{\partial x^1} & \frac{\partial x'^1}{\partial x^2} & \frac{\partial x'^1}{\partial x^3}\\ \frac{\partial x'^2}{\partial x^0} & \frac{\partial x'^2}{\partial x^1} & \frac{\partial x'^2}{\partial x^2} & \frac{\partial x'^2}{\partial x^3}\\ \frac{\partial x'^3}{\partial x^0} & \frac{\partial x'^3}{\partial x^1} & \frac{\partial x'^3}{\partial x^2} & \frac{\partial x'^3}{\partial x^3}\end{vmatrix}. \tag{2.53}$$

For convenience we can also write J as in the first of the following equations:

$$J=\left|\frac{\partial x'}{\partial x}\right|;\quad \left|\frac{\partial x}{\partial x'}\right|=J^{-1}, \tag{2.54}$$

where the second equation, in obvious notation, follows by taking determinants of both sides of the identity

$$(\partial x'^{\mu}/\partial x^{\nu})(\partial x^{\lambda}/\partial x'^{\mu})=\delta^{\lambda}_{\nu}, \tag{2.55}$$

considered as a matrix equation. Equation (2.52) can be written as

$$\mathrm{d}^4x'=J\mathrm{d}^4x. \tag{2.56}$$

With the use of the usual notation $x'^{\mu}{}_{,\nu}\equiv\partial x'^{\mu}/\partial x^{\nu}$, we can write the transformation rule for the covariant metric tensor as follows:

$$g_{\alpha\beta}=x'^{\mu}{}_{,\alpha}g'_{\mu\nu}x'^{\nu}{}_{,\beta}. \tag{2.57}$$

As in (2.55) we consider this as a matrix equation, where in the right hand side the first matrix has its rows specified by α and columns by μ, in the second the rows are given by μ and columns by ν, while in the third matrix the rows and columns are given respectively by ν and β. As before, we denote by g the determinant of the covariant tensor $g_{\alpha\beta}$ considered as a matrix. Taking determinants of both sides of (2.57), we then get

$$g=Jg'J;\quad \text{or } g=J^2g', \tag{2.58}$$

where $g'=\det(g'_{\mu\nu})$. Now g is in general a negative quantity, so we take the square root of the negative of (2.58) to get the following equation:

$$(-g)^{\frac{1}{2}}=J(-g')^{\frac{1}{2}};\quad \varsigma=J\varsigma', \tag{2.59}$$

where in the second equation we have introduced the notation $\varsigma = (-g)^{\frac{1}{2}}$, $\varsigma' = (-g')^{\frac{1}{2}}$, since this quantity occurs in various contexts (the symbol ς is to be read as 'curly g'). Consider now a scalar field quantity which remains invariant under a coordinate transformation. If we call it S, then $S = S'$; S could be $A_\mu B^\mu$, for example, where A_μ is a covariant vector and B^μ a contravariant one. Consider now the following volume integral over some four-dimensional region Ω, and the equations that follow. (There can be no confusion between the Ω used here and the density parameter introduced in Chapter 1.)

$$\int_\Omega S\varsigma \mathrm{d}^4x = \int_\Omega S\varsigma' J \mathrm{d}^4x = \int_{\Omega'} S'\varsigma' \mathrm{d}^4x', \tag{2.60}$$

where Ω' is the region in the coordinates x'^μ that corresponds to Ω, and we have made use of (2.56), (2.59). Equation (2.60) implies that

$$\int_\Omega S\varsigma \mathrm{d}^4x = \text{an invariant}. \tag{2.61}$$

For this reason we call $S\varsigma$ a scalar density, that is, because its volume integral is an invariant. More generally, a set of quantities $\mathcal{F}^\mu{}_\nu$ is said to be a tensor density of rank or weight W if it transforms as follows:

$$\mathcal{F}'^\mu{}_\nu = \left|\frac{\partial x'}{\partial x}\right|^W \frac{\partial x'^\mu}{\partial x^\rho}\frac{\partial x^\sigma}{\partial x'^\nu}\mathcal{F}^\rho{}_\sigma. \tag{2.62}$$

From (2.54) and (2.59) we see that ς is a scalar density of weight -1, so that ς^W is a scalar density of weight $-W$, and hence $\varsigma^W\mathcal{F}^\mu{}_\nu$ is a tensor density of weight zero (when one multiplies two tensor densities, their weights add), that is, it is an ordinary tensor. This can be verified as follows. Let

$$F^\mu{}_\nu = \varsigma^W \mathcal{F}^\mu{}_\nu.$$

Then

$$\begin{aligned} F'^\mu{}_\nu &= (\varsigma')^W \mathcal{F}'^\mu{}_\nu \\ &= (\varsigma^W J^{-W})\left(J^W \frac{\partial x'^\mu}{\partial x^\rho}\frac{\partial x^\sigma}{\partial x'^\nu}\mathcal{F}^\rho{}_\sigma\right) \\ &= \varsigma^W \frac{\partial x'^\mu}{\partial x^\rho}\frac{\partial x^\sigma}{\partial x'^\nu}\mathcal{F}^\rho{}_\sigma = \frac{\partial x'^\mu}{\partial x^\rho}\frac{\partial x^\sigma}{\partial x'^\nu}F^\rho{}_\sigma, \end{aligned} \tag{2.63}$$

which shows that $F^\mu{}_\nu$ is a tensor. Similar results can be obtained for tensors of any kind.

We now introduce the Levi–Civita tensor density $\varepsilon^{\alpha\beta\gamma\delta}$, whose components remain the same in all coordinate systems, namely (we put the coordinates in some definite order such as (t,x,y,z), etc.)

$$\varepsilon^{\alpha\beta\gamma\delta}\begin{cases}=+1, \text{ if } \alpha\beta\gamma\delta \text{ is an even permutation of reference order,}\\ =-1, \text{ if } \alpha\beta\gamma\delta \text{ is an odd permutation of reference order,}\\ = \ \ 0, \text{ if any two or more indices are equal.}\end{cases} \tag{2.64}$$

If we now transform from the coordinate system x^μ to x'^μ, then by definition the new components $\varepsilon'^{\alpha\beta\gamma\delta}$ are given by exactly the same condition as (2.64); on the other hand the two sets of quantities satisfy the following equation:

$$\varepsilon'^{\alpha\beta\gamma\delta}=\left|\frac{\partial x'}{\partial x}\right|^{-1}\frac{\partial x'^{\alpha}}{\partial x^{\lambda}}\frac{\partial x'^{\beta}}{\partial x^{\mu}}\frac{\partial x'^{\gamma}}{\partial x^{\nu}}\frac{\partial x'^{\delta}}{\partial x^{\kappa}}\varepsilon^{\lambda\mu\nu\kappa}. \tag{2.65}$$

This is an identity that follows from the rules for expanding a determinant. But this relation also shows (see (2.62)), that $\varepsilon^{\alpha\beta\gamma\delta}$ is a tensor density of weight -1, so that $\varsigma^{-1}\varepsilon^{\alpha\beta\gamma\delta}$ is an ordinary contravariant tensor. We can form the corresponding covariant tensor density by lowering indices the usual way:

$$\varepsilon_{\alpha\beta\gamma\delta}=g_{\alpha\lambda}g_{\beta\mu}g_{\gamma\nu}g_{\delta\kappa}\varepsilon^{\lambda\mu\nu\kappa}. \tag{2.66}$$

Again making use of expansion of determinants one can show that

$$\varepsilon_{\alpha\beta\gamma\delta}=(-g)\varepsilon^{\alpha\beta\gamma\delta.} \tag{2.67}$$

It can be verified that $\varepsilon_{\alpha\beta\gamma\delta}$ is a covariant tensor density of weight -1. The Levi–Civita tensor density is used for defining the 'dual' of antisymmetric tensors, such as that of the electromagnetic field tensor $F^{\mu\nu}$, the Yang–Mills field tensor, or, with respect to suitable indices, of the Riemann tensor (the latter are needed for some of the so-called 'curvature invariants', which will be mentioned later in connection with singularities).

2.2.3 Gauss and Stokes theorems

We discuss the generalization to curved space of the Gauss or divergence theorem and Stokes theorem, which are used, for example, when one varies a volume or surface integral to derive some field equations. We first write down some relevant identities involving ς (see (2.59)). From its definition we get

$$2\varsigma^{-1}\varsigma'_{,\mu}=g^{-1}g_{,\mu};\quad \varsigma'_{,\nu}=(1/2)\varsigma g^{\lambda\mu}g_{\lambda\mu,\nu}, \tag{2.68}$$

where the second relation can be verified by using the properties of determinants and matrices and the fact that $g^{\lambda\mu}$ is the inverse matrix of $g_{\mu\nu}$. Further, from (2.6a) we see that the covariant divergence $A^{\mu}{}_{;\mu}$ of the contravariant vector A^{μ} is given by

$$A^{\mu}{}_{;\mu} = A^{\mu}{}_{,\mu} + \Gamma^{\mu}_{\nu\mu} A^{\nu} = A^{\mu}{}_{,\mu} + \varsigma^{-1}\varsigma_{,\nu} A^{\nu}, \tag{2.69}$$

where we have used the relation

$$\begin{aligned} \Gamma^{\mu}_{\nu\mu} &= \tfrac{1}{2} g^{\mu\sigma}(g_{\sigma\nu,\mu} + g_{\sigma\mu,\nu} - g_{\nu\mu,\sigma}) \\ &= \tfrac{1}{2} g^{\mu\sigma} g_{\sigma\mu,\nu} = \varsigma^{-1}\varsigma_{,\nu}, \end{aligned} \tag{2.70}$$

the last step following from (2.68). Equation (2.69) then yields, with the use of (2.70), the following relation:

$$\int A^{\mu}{}_{;\mu}\varsigma \mathrm{d}^4x = \int (A^{\mu}\varsigma)_{,\mu} \mathrm{d}^4x. \tag{2.71}$$

From (2.60) we see that the left hand side of (2.71) is an invariant. If the integral is over a finite four-dimensional region Ω, we can use the ordinary divergence theorem to convert (2.71) into a surface integral over the three-dimensional boundary $\partial\Omega$ of the four-dimensional volume Ω. If the covariant divergence of A^{μ} vanishes, we get, with the use of (2.69), a conservation law, as follows:

$$A^{\mu}{}_{;\mu} = 0; \quad (\varsigma A^{\mu})_{,\mu} = 0, \tag{2.72}$$

the second equation being equivalent to the first through (2.69). Integrating the latter equation over a three-dimensional volume V at a definite time x^0, we get

$$\begin{aligned} \left(\int_V \varsigma A^0 \mathrm{d}^3x \right)_{,0} &= -\int_V (\varsigma A^m)_{,m} \mathrm{d}^3x \\ &= (\text{Surface integral over } \partial V, \text{ boundary of } V). \end{aligned} \tag{2.73}$$

This relation can be looked upon as the conservation of a fluid whose density (we are here using 'density' in the usual sense) is ςA^0 and whose motion is determined by the three-dimensional vector $\varsigma A^m (m = 1,2,3)$. If there is no flow across the boundary, (2.73) shows that

$$\int \varsigma A^0 \mathrm{d}^3x = \text{constant}.$$

This example illustrates the circumstance that when considering volume and surface integrals and conservation laws, it is the tensor density (vector

density in this case) ςA^{μ} that is more relevant. However, these results are not in general applicable, at least not in the above form, for a tensor with more than one suffix. Also, unlike the case of a scalar density (2.61), the integral

$$\int T^{\mu\nu}\varsigma \mathrm{d}^4x$$

is not in general a tensor, because the integral gives essentially sums of terms at different points, which transform differently, so the sum or integral does not transform in any simple manner. In the special case of an antisymmetric tensor $F^{\mu\nu}=-F^{\mu\nu}$, a conservation law can be obtained as follows. We have

$$F^{\mu\nu}{}_{;\sigma}=F^{\mu\nu}{}_{,\sigma}+\Gamma^{\mu}_{\sigma\rho}F^{\rho\nu}+\Gamma^{\nu}_{\sigma\rho}F^{\mu\rho},$$

whence

$$\begin{aligned}F^{\mu\nu}{}_{;\nu}&=F^{\mu\nu}{}_{,\nu}+\Gamma^{\mu}_{\nu\rho}F^{\rho\nu}+\Gamma^{\nu}_{\nu\rho}F^{\mu\rho}\\&=F^{\mu\nu}{}_{,\nu}+\varsigma^{-1}\varsigma_{,\rho}F^{\mu\rho},\end{aligned}\tag{2.74}$$

where we have used (2.70) and the fact that $\Gamma^{\mu}_{\nu\rho}F^{\nu\rho}$ vanishes (because $\Gamma^{\mu}_{\nu\rho}$ is symmetric while $F^{\rho\nu}$ is antisymmetric in ρ and ν). From (2.74) we get

$$\varsigma F^{\mu\nu}{}_{;\nu}=(\varsigma F^{\mu\nu})_{,\nu}.\tag{2.75}$$

With the use of reasoning similar to that used in (2.72) and (2.73), we see from (2.75) that

$$\int \varsigma F^{\mu\nu}{}_{;\nu}\mathrm{d}^4x=\text{surface integral},$$

from which a conservation law follows.

For a symmetric tensor $Y^{\mu\nu}=Y^{\nu\mu}$ we can get a conservation law with an additional term. In this case

$$Y_{\mu}{}^{\nu}{}_{;\sigma}=Y_{\mu}{}^{\nu}{}_{,\sigma}-\Gamma^{\alpha}_{\mu\sigma}Y_{\alpha}{}^{\nu}+\Gamma^{\nu}_{\sigma\alpha}Y_{\mu}{}^{\alpha}.$$

We set $\nu=\sigma$ and use (2.70) to get

$$Y_{\mu}{}^{\nu}{}_{;\nu}=Y_{\mu}{}^{\nu}{}_{,\nu}+\varsigma^{-1}\varsigma_{,\alpha}Y_{\mu}{}^{\alpha}-\Gamma^{\alpha}_{\mu\nu}Y_{\alpha}{}^{\nu}.\tag{2.76}$$

We can transform the last term as follows:

$$\begin{aligned}\Gamma^{\alpha}_{\mu\nu}Y_{\alpha}{}^{\nu}&=g_{\beta\alpha}\Gamma^{\alpha}_{\mu\nu}Y^{\beta\nu}=g_{\nu\alpha}\Gamma^{\alpha}_{\mu\beta}Y^{\nu\beta}=g_{\nu\alpha}Y^{\beta\nu}\Gamma^{\alpha}_{\mu\beta}\\&=(1/2)(g_{\beta\alpha}\Gamma^{\alpha}_{\mu\nu}+g_{\nu\alpha}\Gamma^{\alpha}_{\mu\beta})Y^{\beta\nu}=(1/2)g_{\beta\nu,\mu}Y^{\beta\nu}.\end{aligned}$$

Using this in (2.76) we get

$$\varsigma Y_{\mu}{}^{\nu}{}_{;\nu} = (\varsigma Y_{\mu}{}^{\nu})_{,\nu} - \tfrac{1}{2}\varsigma g_{\beta\nu,\mu} Y^{\beta\nu}. \tag{2.77}$$

If we now have $Y_{\mu}{}^{\nu}{}_{;\nu} = 0$, we integrate (2.77) over a three dimensional volume V at time x^0, to obtain the following result:

$$\int_V (\varsigma Y_{\mu}{}^{0}\mathrm{d}^3x)_{,0} = -\int_V (\varsigma Y_{\mu}{}^{\mu})_{,m}\mathrm{d}^3x + \tfrac{1}{2}\int_V \varsigma g_{\beta\nu,\mu} Y^{\beta\nu}\mathrm{d}^3x$$

$$= (\text{Integral over surface } \partial V \text{ of } V) + \tfrac{1}{2}\varsigma \int g_{\beta\nu,\mu} Y^{\beta\nu}\mathrm{d}^3x. \tag{2.78}$$

Even if the surface integral vanishes, the quantities

$$\theta_{\mu} = \int \varsigma Y_{\mu}{}^{0}\mathrm{d}^3x$$

cannot be considered as constant because of the last term on the right hand side of (2.78), which could represent the generation or disappearance of some quantities in the volume V; in other words the volume V has a 'source' or a 'sink' for some physical quantity. Such a situation arises, for example, for the energy–momentum tensor that occurs on the right hand side of Einstein's equations (2.22). This is a symmetric tensor, and the covariant divergence of its contravariant form $T^{\mu\nu}$ vanishes (see (2.25)). The reason one gets an 'additional term' here is that the energy momentum of matter has to be balanced by that of the gravitational field, which is not easy to define in a coordinate independent manner (see Landau and Lifshitz 1975, p. 280; Dirac 1975, p. 64; Weinberg 1972, p. 165).

We consider now Stokes's theorem. From (2.6b) it is readily verified that the covariant curl equals the ordinary curl:

$$A_{\mu;\nu} - A_{\nu;\mu} = A_{\mu,\nu} - A_{\nu,\mu}. \tag{2.79}$$

This does not in general hold for a contravariant vector. Put $\mu = 1$, $\nu = 2$ to get

$$A_{1;2} - A_{2;1} = A_{1,2} - A_{2,1}. \tag{2.80}$$

Integrating this over an area of a surface S given by $x^0 = \text{constant}$, $x^3 = \text{constant}$, and using the ordinary form of Stokes's theorem, we get

$$\iint_S (A_{1;2} - A_{2;1})\mathrm{d}x^1\mathrm{d}x^2 = \iint_S (A_{1,2} - A_{2,1})\mathrm{d}x^1\mathrm{d}x^2$$

$$= \int_{\partial S} (A_1\mathrm{d}x^1 + A_2\mathrm{d}x^2), \tag{2.81}$$

where the integral at the end is over the perimeter ∂S of the area S. We will express this in an invariant manner. An element of surface $\mathrm{d}S^{\mu\nu}$ given by two infinitesimal contravariant vectors ξ^μ and ζ^μ is given by

$$\mathrm{d}S^{\mu\nu} = \xi^\mu\zeta^\nu - \xi^\nu\zeta^\mu. \tag{2.82}$$

For example, if $\xi^\mu = (0,\mathrm{d}x^1,0,0)$, $\zeta^\mu = (0,0,\mathrm{d}x^2,0)$, then

$$\mathrm{d}S^{12} = \mathrm{d}x^1\mathrm{d}x^2,\ \mathrm{d}S^{21} = -\mathrm{d}x^1\mathrm{d}x^2,$$

the other components being zero. Thus (2.81) becomes

$$\tfrac{1}{2}\iint_{\text{surface}} (A_{\mu;\nu} - A_{\nu;\mu})\mathrm{d}S^{\mu\nu} = \int_{\text{perimeter}} A_\mu \mathrm{d}x^\mu. \tag{2.83}$$

In this form Stokes's theorem can be used for curved (Riemannian) spaces.

2.2.4 The action principle for gravitation

Consider the quantity

$$I = \int_\Omega \varsigma R \mathrm{d}^4x, \tag{2.84}$$

where Ω is a given four-dimensional region. From (2.60) we see that I is a scalar (invariant) quantity. We will consider the variation of the quantity δI, when the $g_{\mu\nu}$ are varied by an infinitesimal amount: $g_{\mu\nu} \to g_{\mu\nu} + \delta g_{\mu\nu}$, such that the variations vanish on the boundary $\partial\Omega$ of Ω. If we put $\delta I = 0$, we will obtain Einstein's vacuum field equations:

$$R_{\mu\nu} = 0. \tag{2.85}$$

From (2.71) and (2.20) we get

$$R = g^{\mu\nu}R_{\mu\nu} = g^{\mu\nu}(\Gamma^\lambda_{\mu\nu,\lambda} - \Gamma^\lambda_{\mu\lambda,\nu} + \Gamma^\lambda_{\mu\nu}\Gamma^\sigma_{\lambda\sigma} - \Gamma^\sigma_{\mu\lambda}\Gamma^\lambda_{\nu\sigma}) = R^* + \mathrm{Q}, \tag{2.86a}$$

where

$$R^* = g^{\mu\nu}(\Gamma^\lambda_{\mu\nu,\lambda} - \Gamma^\lambda_{\mu\lambda,\nu}),\quad Q = g^{\mu\nu}(\Gamma^\lambda_{\mu\nu}\Gamma^\sigma_{\lambda\sigma} - \Gamma^\sigma_{\mu\lambda}\Gamma^\lambda_{\nu\sigma}). \tag{2.86b}$$

We first remove the second derivatives of the $g_{\mu\nu}$ from I given by (2.84); these occur in the expression R^*. We get

$$\varsigma R^* = -(\varsigma g^{\mu\nu}\Gamma^\sigma_{\mu\sigma})_{,\nu} + (\varsigma g^{\mu\nu}\Gamma^\sigma_{\mu\nu})_{,\sigma} + (\varsigma g^{\mu\nu})_{,\nu}\Gamma^\sigma_{\mu\sigma} - (\varsigma g^{\mu\nu})_{,\sigma}\Gamma^\sigma_{\mu\nu}. \tag{2.87}$$

We can use the divergence theorem to convert the first two integrals into surface integrals over $\partial\Omega$, and so they will not contribute to the variation δI since the $\delta g_{\mu\nu}$ vanish on $\partial\Omega$. With the use of (2.7) one can show that

$$(\varsigma g^{\mu\nu})_{,\sigma} = (-g^{\nu\beta}\Gamma^{\mu}_{\beta\sigma} - g^{\mu\alpha}\Gamma^{\nu}_{\alpha\sigma} + g^{\mu\nu}\Gamma^{\rho}_{\sigma\rho})\varsigma. \tag{2.88}$$

Setting $\nu = \sigma$ we get (since the second and third terms on the right cancel):

$$(g^{\mu\nu}\varsigma)_{,\nu} = -g^{\nu\beta}\Gamma^{\mu}_{\beta\nu}\varsigma. \tag{2.89}$$

With the use of (2.88), (2.89), the last two terms of (2.87) become

$$\begin{aligned} &-g^{\nu\beta}\Gamma^{\mu}_{\beta\nu}\Gamma^{\sigma}_{\mu\sigma}\varsigma + (g^{\nu\beta}\Gamma^{\mu}_{\beta\sigma} + g^{\mu\alpha}\Gamma^{\nu}_{\alpha\sigma} - g^{\mu\nu}\Gamma^{\beta}_{\sigma\beta})\varsigma\Gamma^{\sigma}_{\mu\nu} \\ &= \varsigma[-g^{\nu\beta}\Gamma^{\mu}_{\beta\nu}\Gamma^{\sigma}_{\mu\sigma} + (2g^{\nu\beta}\Gamma^{\mu}_{\beta\sigma} - g^{\mu\nu}\Gamma^{\beta}_{\sigma\beta})\Gamma^{\sigma}_{\mu\nu}] \\ &= \varsigma[-2g^{\nu\beta}\Gamma^{\mu}_{\beta\nu}\Gamma^{\sigma}_{\mu\sigma} + 2g^{\nu\beta}\Gamma^{\mu}_{\beta\sigma}\Gamma^{\sigma}_{\mu\nu}] = -2\varsigma Q. \end{aligned} \tag{2.90}$$

Thus

$$I = -\int \varsigma Q \mathrm{d}^4x. \tag{2.91}$$

Although the integrand is not a scalar quantity, it is convenient for the present purpose, since it contains only $g_{\mu\nu}$ and their first derivatives, being homogeneous of the second degree in the derivatives.

In dynamical problems the action I is in fact the time integral of the Lagrangian L, so that the latter is given by

$$L = \int \mathcal{L}\mathrm{d}x'\mathrm{d}x^2\mathrm{d}x^3,$$

where $\mathcal{L}$ is the Lagrangian density, and the action I can be taken as a time integral of L, and a space-time integral of $\mathcal{L}$, as follows:

$$I = \int L\mathrm{d}x^0 = \int \mathcal{L}\mathrm{d}^4x; \quad \mathcal{L} = \varsigma R. \tag{2.92}$$

The $g_{\mu\nu}$ can be considered as coordinates and their time derivatives as velocities. Thus, as in ordinary dynamics, the Lagrangian is a non-homogeneous quadratic in the velocities.

We consider the variation of the two parts of ςQ (see (2.86b)), as follows:

$$\begin{aligned} \delta(\Gamma^{\alpha}_{\mu\nu}\Gamma^{\beta}_{\alpha\beta}g^{\mu\nu}\varsigma) &= \Gamma^{\alpha}_{\mu\nu}\delta(\Gamma^{\beta}_{\alpha\beta}g^{\mu\nu}\varsigma) + \Gamma^{\beta}_{\alpha\beta}g^{\mu\nu}\varsigma\delta\Gamma^{\alpha}_{\mu\nu} \\ &= \Gamma^{\alpha}_{\mu\nu}\delta(g^{\mu\nu}\varsigma_{,\alpha}) + \Gamma^{\beta}_{\alpha\beta}\{\delta(\Gamma^{\alpha}_{\mu\nu}g^{\mu\nu}\varsigma) - \Gamma^{\alpha}_{\mu\nu}\delta(g^{\mu\nu}\varsigma)\} \\ &= \Gamma^{\alpha}_{\mu\nu}\delta(g^{\mu\nu}\varsigma_{,\alpha}) - \Gamma^{\beta}_{\alpha\beta}\delta(g^{\alpha\nu}\varsigma)_{,\nu} - \Gamma^{\beta}_{\alpha\beta}\Gamma^{\alpha}_{\mu\nu}\delta(g^{\mu\nu}\varsigma), \end{aligned} \tag{2.93a}$$

$$\begin{aligned}\delta(\Gamma^{\beta}_{\mu\alpha}\Gamma^{\alpha}_{\nu\beta}g^{\mu\nu}\varsigma) &= 2(\delta\Gamma^{\beta}_{\mu\alpha})\Gamma^{\alpha}_{\nu\beta}g^{\mu\nu}\varsigma + \Gamma^{\beta}_{\mu\alpha}\Gamma^{\alpha}_{\nu\beta}\delta(g^{\mu\nu}\varsigma)\\ &= 2\delta(\Gamma^{\beta}_{\mu\alpha}g^{\mu\nu}\varsigma)\Gamma^{\alpha}_{\nu\beta} - 2\Gamma^{\beta}_{\mu\alpha}\delta(g^{\mu\nu}\varsigma)\Gamma^{\alpha}_{\nu\beta} + \Gamma^{\beta}_{\mu\alpha}\Gamma^{\alpha}_{\nu\beta}\delta(g^{\mu\nu}\varsigma)\\ &= 2\delta(\Gamma^{\beta}_{\mu\alpha}g^{\mu\nu}\varsigma)\Gamma^{\alpha}_{\nu\beta} - \Gamma^{\beta}_{\mu\alpha}\Gamma^{\alpha}_{\nu\beta}\delta(g^{\mu\nu}\varsigma)\\ &= -\delta(g^{\nu\beta}{}_{,\alpha}\varsigma)\Gamma^{\alpha}_{\nu\beta} - \Gamma^{\beta}_{\mu\alpha}\Gamma^{\alpha}_{\nu\beta}\delta(g^{\mu\nu}\varsigma). \qquad (2.93b)\end{aligned}$$

Subtracting,we get

$$\begin{aligned}\delta(Q\varsigma) &= \Gamma^{\alpha}_{\mu\nu}\delta(g^{\mu\nu}\varsigma)_{,\alpha} - \Gamma^{\beta}_{\alpha\beta}\delta(g^{\alpha\nu}\varsigma)_{,\nu}\\ &\quad + (\Gamma^{\beta}_{\mu\alpha}\Gamma^{\alpha}_{\nu\beta} - \Gamma^{\beta}_{\alpha\beta}\Gamma^{\alpha}_{\mu\nu})\delta(g^{\mu\nu}\varsigma)\\ &= [\Gamma^{\alpha}_{\mu\nu}\delta(g^{\mu\nu}\varsigma)]_{,\alpha} - [\Gamma^{\beta}_{\alpha\beta}\delta(g^{\alpha\nu}\varsigma)]_{,\nu}\\ &\quad + \{-\Gamma^{\alpha}_{\mu\nu,\alpha} + \Gamma^{\beta}_{\mu\beta,\nu} + \Gamma^{\beta}_{\mu\alpha}\Gamma^{\alpha}_{\nu\beta} - \Gamma^{\beta}_{\alpha\beta}\Gamma^{\alpha}_{\mu\nu}\}\delta(g^{\mu\nu}\varsigma). \qquad (2.94)\end{aligned}$$

The first two terms, being perfect differentials, may be transformed, as usual, using the divergence theorem to surface integrals, which vanish because the variations vanish on the surface. The expression in the curly brackets is just $R_{\mu\nu}$. Thus (see (2.91))

$$\delta I = \delta\int \mathcal{L}\mathrm{d}^4x = -\int R_{\mu\nu}\delta(g^{\mu\nu}\varsigma)\mathrm{d}^4x. \qquad (2.95)$$

Since the $\delta g_{\mu\nu}$ are arbitrary, the quantities $\delta(g^{\mu\nu}\varsigma)$ are also arbitrary, and hence $\delta I = 0$ implies the vacuum Einstein equations (2.85).

By taking variation of (2.4) one can readily show that

$$\delta g^{\mu\nu} = -g^{\mu\alpha}g^{\nu\beta}\delta g_{\alpha\beta}.$$

With the use of (2.68) one can obtain the following relation:

$$\delta\varsigma = \tfrac{1}{2}\varsigma g^{\alpha\beta}\delta g_{\alpha\beta},$$

so that

$$\delta(g^{\mu\nu}\varsigma) = -(g^{\mu\alpha}g^{\nu\beta} - \tfrac{1}{2}g^{\mu\nu}g^{\alpha\beta})\varsigma\delta g_{\alpha\beta}. \qquad (2.96)$$

Thus (2.95) can be written as

$$\delta I = -\int (R^{\mu\nu} - \tfrac{1}{2}g^{\mu\nu}R)\varsigma\delta g_{\mu\nu}\mathrm{d}^4x, \qquad (2.97)$$

leading to

$$R^{\mu\nu} - \tfrac{1}{2}g^{\mu\nu}R = 0, \qquad (2.98)$$

which is another form of the vacuum Einstein equations (see (2.22)).

The geodesic equation (2.26) can also be obtained by a variation principle, if it is not a null geodesic. Consider the quantity

$$S_{AB} = \int_A^B \mathrm{d}s, \tag{2.99}$$

which represents the 'length' (in the case of time-like curves this is the proper time) from the point A to the point B of the curve. Let each point of the curve with coordinate x^μ be moved to $x^\mu + \mathrm{d}x^\mu$. If $\mathrm{d}x^\mu$ is an element along the curve, we have

$$\mathrm{d}s^2 = g_{\mu\nu}\mathrm{d}x^\mu \mathrm{d}x^\nu.$$

Taking variations of both sides, we get

$$2\mathrm{d}s\delta(\mathrm{d}s) = \mathrm{d}x^\mu \mathrm{d}x^\nu g_{\mu\nu,\lambda}\delta x^\lambda + 2g_{\mu\lambda}\mathrm{d}x^\mu \delta(\mathrm{d}x^\lambda). \tag{2.100}$$

Further, with $u^\mu \mathrm{d}s = \mathrm{d}x^\mu$ (see (2.24)), $\delta(\mathrm{d}x^\lambda) = \mathrm{d}(\delta x^\lambda)$, we get from (2.100) the following expression for $\delta(\mathrm{d}s)$:

$$\delta(\mathrm{d}s) = \left\{ \tfrac{1}{2} g_{\mu\nu,\lambda} \frac{\mathrm{d}x^\mu}{\mathrm{d}s} \frac{\mathrm{d}x^\nu}{\mathrm{d}s} \delta x^\lambda + g_{\mu\lambda} \frac{\mathrm{d}x^\mu}{\mathrm{d}s} \frac{\mathrm{d}\delta x^\lambda}{\mathrm{d}s} \right\} \mathrm{d}s.$$

Therefore,

$$\delta \int_A^B \mathrm{d}s = \int_A^B \delta(\mathrm{d}s) = \int_A^B \left\{ \tfrac{1}{2} g_{\mu\nu,\lambda} u^\mu u^\nu \delta x^\lambda + g_{\mu\lambda} u^\mu \frac{\mathrm{d}}{\mathrm{d}s}(\delta x^\lambda) \right\} \mathrm{d}s. \tag{2.101}$$

We carry out partial integration with respect to s and use the fact that $\delta x^\lambda = 0$ at A and B, to get

$$\delta S_{AB} = \delta \int_A^B \mathrm{d}s = \int_A^B \left\{ \tfrac{1}{2} g_{\mu\nu,\lambda} u^\mu u^\nu - \frac{\mathrm{d}}{\mathrm{d}s}(g_{\mu\lambda} u^\mu) \right\} \delta x^\lambda \mathrm{d}s. \tag{2.102}$$

Since the δx^λ are arbitrary, for $\delta S_{AB} = 0$, we get

$$\frac{\mathrm{d}}{\mathrm{d}s}(g_{\mu\lambda} u^\mu) - \tfrac{1}{2} g_{\mu\nu,\lambda} u^\mu u^\nu = 0. \tag{2.103}$$

The first term can be transformed as follows:

$$\frac{\mathrm{d}}{\mathrm{d}s}(g_{\mu\lambda} u^\mu) = g_{\mu\lambda} \frac{\mathrm{d}u^\mu}{\mathrm{d}s} + g_{\mu\lambda,\nu} \frac{\mathrm{d}x^\nu}{\mathrm{d}s} u^\mu$$

$$= g_{\mu\lambda} \frac{\mathrm{d}u^\mu}{\mathrm{d}s} + \tfrac{1}{2}(g_{\mu\lambda,\nu} + g_{\nu\lambda,\mu}) u^\mu u^\nu.$$

Substituting this in (2.103), we get

$$g_{\mu\lambda}\frac{\mathrm{d}u^{\mu}}{\mathrm{d}s}+\tfrac{1}{2}(g_{\mu\lambda,\nu}+g_{\nu\lambda,\mu}-g_{\mu\nu,\lambda})u^{\mu}u^{\nu}=0.$$

Multiplying by $g^{\lambda\sigma}$ and using (2.7), we finally obtain the following relation:

$$\frac{\mathrm{d}u^{\sigma}}{\mathrm{d}s}+\Gamma^{\sigma}_{\mu\nu}u^{\mu}u^{\nu}=0, \tag{2.104}$$

which is the geodesic equation (2.26) written in terms of u^{μ}.

2.2.5 Some further topics

In this section we consider some additional topics. First we consider the action principle in the presence of matter. Before we do this, we will consider a description of the interior of matter that leads, for example, to the energy–momentum tensor given by (2.23). We will deal with the simpler situation in which the pressure p vanishes and the material particles move along geodesics. We have in mind a distribution of matter in which the velocity varies from one element to a neighbouring one continuously. The worldlines of material particles fill up all space-time or a portion of it, a typical worldline being denoted by $z^{\mu}(s)$, in which the points along the worldline are distinguished by values of a parameter s which measures the interval along the line. The interval $\mathrm{d}s$ between points z^{μ} and $z^{\mu}+\mathrm{d}z^{\mu}$ satisfies

$$\mathrm{d}s^2=g_{\mu\nu}\mathrm{d}z^{\mu}\mathrm{d}z^{\nu}, \tag{2.105}$$

so that the four-velocity, given by the first equation in the following, satisfies the second equation (see (2.24)):

$$u^{\mu}=\mathrm{d}z^{\mu}/\mathrm{d}s; \quad g_{\mu\nu}u^{\mu}u^{\nu}=1. \tag{2.106}$$

The four-velocity u^{μ} can be considered as a contravariant vector field whose components are functions of the space-time point x^{μ} ($\mu=0,1,2,3$), sometimes written as x, as before. There is a unique worldline passing through each space-time point x, and they may all be indentified by the point ξ^{μ} at which they intersect some space-like hypersurface given by

$$f(\xi^0,\xi^1,\xi^2,\xi^3)=0. \tag{2.107}$$

A one-to-one correspondence can be set up between points on this three-dimensional space-like hypersurface and the set of worldlines filling all of

space-time. If the surface $f=0$ is designated by $\xi^0=0$, then the worldlines can be identified by the point $\boldsymbol{\xi}=(\xi^1,\xi^2,\xi^3)$ on $\xi^0=0$ through which it passes, and the coordinates on a typical such worldline can be written as:

$$z^\mu(s;\boldsymbol{\xi}), \text{ with } z^\mu(0;\boldsymbol{\xi})=(0,\xi^i). \tag{2.108}$$

However, having set up this coordinate system for the worldlines, or the flow of matter, we will simply assume, as mentioned, that the four-velocity vector u^μ is a function of the position (or 'event') coordinates x^μ.

Taking the covariant derivative of the second relation in (2.106), which is the same as the ordinary derivative for a scalar, we get

$$0=(g_{\mu\nu}u^\mu u^\nu)_{,\sigma}=(g_{\mu\nu}u^\mu u^\nu)_{;\sigma}=g_{\mu\nu}(u^\mu{}_{;\sigma}u^\nu+u^\mu u^\nu{}_{;\sigma}), \tag{2.109}$$

where we have used the fact that the covariant derivative of $g_{\mu\nu}$ vanishes, and the symmetry of $g_{\mu\nu}$. From (2.109) we get

$$u_\nu u^\nu{}_{;\sigma}=0. \tag{2.110}$$

Just like the fact that the charge density ρ and the current $j^m=(j^1,j^2,j^3)$ form a four-vector current J^μ, with

$$J^0=\rho,\ J^i=j^i,\ (i=1,2,3), \tag{2.111}$$

in electromagnetic theory, so we can define a scalar field ρ and the corresponding vector field ρu^μ which determine the density and flow of matter. We have seen (see (2.73) and following discussion) that the ordinary density of matter is not a component of a four-vector, but that of a four-vector density, which is obtained by multiplying the four-vector by $\varsigma=(-g)^{\frac{1}{2}}$. Thus the density here is given by $\varsigma\rho u^0$ and the flow or the current by $\varsigma\rho u^i$ $(i=1,2,3)$. The equation for the conservation of matter (equation of continuity) is:

$$(\varsigma\rho u^\mu)_{,\mu}=0,$$

which implies (see (2.69) and (2.72))

$$(\rho u^\mu)_{;\mu}=0. \tag{2.112}$$

The matter under consideration has energy density $(\rho u^0)u^0\varsigma$ and energy flux $(\rho u^0)u^i\varsigma$. Similarly, there will be a momentum density $(\rho u^0)u^n\varsigma$ and a momentum flux $\rho u^n u^m\varsigma$. These properties are reflected in the tensor (this discussion is taken from Dirac 1975, 1996, p. 45):

$$T^{\mu\nu}=\rho u^\mu u^\nu, \tag{2.113}$$

with $\varsigma T^{\mu\nu}$ giving the density and flux of energy and momentum. The symmetric tensor $T^{\mu\nu}$ is the material energy–momentum tensor. From (2.112), (2.113) we get

$$\begin{aligned} T^{\mu\nu}{}_{;\nu} &= (\rho u^\mu u^\nu)_{;\nu} \\ &= (\rho u^\nu)_{;\nu} u^\mu + \rho u^\nu u^\mu{}_{;\nu} \\ &= \rho u^\nu u^\mu{}_{;\nu}. \end{aligned} \tag{2.114}$$

If, as mentioned earlier, u^ν is regarded as a field function (i.e., meaningful not just on one worldline but a whole set of worldlines filling up all spacetime or a region thereof), we obtain the following relations:

$$\mathrm{d}u^\mu/\mathrm{d}s = (\partial u^\mu/\partial x^\nu)(\mathrm{d}x^\nu/\mathrm{d}s) = u^\mu{}_{,\nu} u^\nu, \tag{2.115}$$

whence we get, using (2.104),

$$\begin{aligned} (u^\mu{}_{,\nu} u^\nu + \Gamma^\mu_{\nu\sigma} u^\nu u^\sigma) &= (u^\mu{}_{,\nu} + \Gamma^\mu_{\nu\sigma} u^\sigma) u^\nu \\ &= u^\mu{}_{;\nu} u^\nu = 0. \end{aligned} \tag{2.116}$$

With the use of (2.114) and (2.116) one can get the following relation:

$$T^{\mu\nu}{}_{;\nu} = 0, \tag{2.117}$$

so that the tensor $T^{\mu\nu}$ defined by (2.113) can be used on the right hand side of Einstein's equations (2.22). However, the tensor given by (2.113) is a special case of that which occurs in (2.23), being obtained from the latter by setting $p = 0$. This zero-pressure case obtains when there is no random motion of the material particles that is associated with pressure, so that the particles move solely under the influence of gravitation and so move along geodesics given by (2.104), leading to (2.116). This zero-pressure form of matter is usually referred to as 'dust', and arises in various situations including cosmological ones, as we shall see later.

We will continue a little further the derivation of Einstein's equations in the case of dust to introduce the Newtonian approximation and clarify certain minor issues. From the property (2.21) of the Einstein tensor $R^{\mu\nu} - \frac{1}{2} g^{\mu\nu} R$ and from (2.113), (2.117) we can set

$$R^{\mu\nu} - \tfrac{1}{2} g^{\mu\nu} R = k T^{\mu\nu} = k \rho u^\mu u^\nu. \tag{2.118}$$

(To emphasize the difference in the zero-pressure case and the non-zero pressure case, we will use ρ for the density in the former case as at present, but continue to use ε for the mass-energy density in the non-zero pressure case as in (2.23).) To find the constant k, we have to resort to the Newtonian approximation, for which we consider the *static* metric, the

components of which are independent of the time, and the 'mixed' components are zero:

$$g_{\mu\nu,0} = 0,\ g_{i0} = 0,\ i = 1,2,3, \tag{2.119a}$$

whence we get

$$g^{i0} = 0,\ i = 1,2,3;\ g^{00} = (g_{00})^{-1}. \tag{2.119b}$$

With the use of (2.7), we readily see that (2.119a,b) imply

$$\Gamma^{i}_{0j} = 0,\ i,j = 1,2,3. \tag{2.120}$$

Because of (2.119a) we can write the second equation in (2.106) as follows:

$$g_{00}(u^0)^2 + g_{ij}u^i u^j = 1. \tag{2.121}$$

For particles moving slowly with respect to the speed of light, the second set of terms on the left hand side is small compared to the first term (since the u^i are of the order of v/c, where v is a typical velocity), so we get

$$g_{00}(u^0)^2 = 1. \tag{2.122}$$

If the particle moves along a geodesic, one obtains from (2.104), neglecting second order quantities (that is, terms proportional to $(v/c)^2$, etc.),

$$\mathrm{d}u^i/\mathrm{d}s = -\Gamma^{i}_{00}(u^0)^2 = (1/2)g^{ij}g_{00,j}(u^0)^2. \tag{2.123a}$$

To first order we also have

$$\mathrm{d}u^i/\mathrm{d}s = (\partial u^i/\partial x^\mu)(\mathrm{d}x^\mu/\mathrm{d}s) = (\partial u^i/\partial x^0)u^0. \tag{2.123b}$$

Equating the right hand sides of (2.123a,b) and cancelling a factor u^0, results in (from (2.122) $u^0 = (g_{00})^{-\frac{1}{2}}$)

$$(\partial u^i/\partial x^0) = (1/2)g^{ij}g_{00,j}u^0 = g^{ij}(g_{00}^{\frac{1}{2}})_{,j}. \tag{2.124}$$

With the use of (2.119a) and (2.124) we get

$$\begin{aligned} g_{ik}(\partial u^k/\partial x^0) &= g_{ik}g^{kj}(g_{00}^{\frac{1}{2}})_{,j} \\ &= \delta^j_i(g_{00}^{\frac{1}{2}})_{,j} = (g_{00}^{\frac{1}{2}})_{,i} = (\partial u_i/\partial x^0). \end{aligned} \tag{2.125}$$

This equation is analogous to the Newtonian equation of motion, in that the 'acceleration' $\partial u_i/\partial x^0$, in units employed here, is equal to the gradient of a scalar, in this case $g_{00}^{\frac{1}{2}}$, which plays the role of the Newtonian gravitational potential. Assuming the gravitational field to be weak, the $g_{\mu\nu}$ can be taken to be constant (i.e., independent of time, as in (2.119a)), and the

$g_{\mu\nu,\sigma}$ and hence all Christoffel symbols to be small. Under these conditions the vacuum Einstein equations $R_{\mu\nu}=0$ become (neglecting products of Γs)

$$\Gamma^{\alpha}_{\mu\alpha,\nu}-\Gamma^{\alpha}_{\mu\nu,\alpha}=0. \tag{2.126}$$

This is equivalent to the following equation:

$$g^{\rho\sigma}(g_{\rho\sigma,\mu\nu}-g_{\nu\sigma,\mu\rho}-g_{\mu\rho,\nu\sigma}+g_{\mu\nu,\rho\sigma})=0. \tag{2.127}$$

If we set $\mu=\nu=0$ and $g_{\mu\nu,0}=0$, Equation (2.127) reduces to

$$g^{mn}g_{00,mn}=0. \tag{2.128}$$

This is analogous to Laplace's equation. If we choose units so that g_{00} is approximately unity, we may take

$$g_{00}=1+2V/c^2, \tag{2.129}$$

so that $g_{00}^{\frac{1}{2}}\cong 1+V/c^2$, and from (2.125) we see that V may be identified with the Newtonian gravitational potential (see below).

Going back to (2.118), by multiplying by $g^{\mu\nu}$ we get (introducing c):

$$-R=c^2k\rho,$$

so that (2.118) becomes

$$R_{\mu\nu}=c^2k\rho(u_\mu u_\nu-(1/2)g_{\mu\nu}). \tag{2.130}$$

In the weak field approximation which yields (2.127), we now get

$$(1/2)g^{\rho\sigma}(g_{\rho\sigma,\mu\nu}-g_{\nu\sigma,\mu\rho}-g_{\mu\rho,\nu\sigma}+g_{\mu\nu,\rho\sigma})=kc^2\rho(u_\mu u_\nu-\tfrac{1}{2}g_{\mu\nu}). \tag{2.131}$$

Consider again a static field produced by a static (not moving) distribution of matter, so that $u_0=1$, $u_i=0$. With $\mu=\nu=0$, in (2.131), one gets

$$(1/2)g^{mn}g_{00,mn}=(1/2)c^2k\rho(1-\tfrac{1}{2}g_{00}). \tag{2.132}$$

If we substitute $g_{00}=1+2V/c^2$, and keep the leading terms in powers of c we see that g^{mn} may be taken as δ^{mn} (the Kronecker delta), and g_{00} on the right hand side may be taken as unity. This yields the following equation:

$$\nabla^2V=(1/2)c^4k\rho=4\pi G\rho, \tag{2.133}$$

the last relation coming from Poisson's equation for V, G being Newton's gravitational constant (6.67×10^{-8} $\mathrm{cm^3g^{-1}s^{-2}}$). In (2.133) we have used $g^{mn}V_{,mn}\cong\delta^{mn}V_{,mn}=(\partial^2/(\partial x^1)^2+\partial^2/(\partial x^2)^2+\partial^2/(\partial x^3)^2)V\equiv\nabla^2V$. From (2.133) we get $k=8\pi G/c^4$. This is used in (2.22). (The calculations of this section follow closely those of sections 16 and 25 of Dirac 1975, 1996.)

The derivation of the Newtonian approximation could have been shortened, but the longer discussion given here touches on points of somewhat wider interest.

3
The Robertson–Walker metric

3.1 A simple derivation of the Robertson–Walker metric

As we saw in the first chapter, the universe appears to be homogeneous and isotropic around us on scales of more than a 100 million light years or so, so that on this scale the density of galaxies is approximately the same and all directions from us appear to be equivalent. From these observations one is led to the Cosmological Principle which states that the universe looks the same from all positions in space at a particular time, and that all directions in space at any point are equivalent. This is an intuitive statement of the Cosmological Principle which needs to be made more precise. For example, what does one mean by 'a particular time'? In Newtonian physics this concept is unambiguous. In special relativity the concept becomes well-defined if one chooses a particular inertial frame. In general relativity, however, there are no global inertial frames. To define 'a moment of time' in general relativity which is valid globally, a particular set of circumstances are necessary, which, in fact, are satisfied by a homogeneous and isotropic universe.

To define 'a particular time' in general relativity which is valid globally in this case, we proceed as follows. Introduce a series of non-intersecting space-like hypersurfaces, that is, surfaces any two points of which can be connected to each other by a curve lying entirely in the hypersurface which is space-like everywhere. We make the assumption that all galaxies lie on such a hypersurface in such a manner that the surface of simultaneity of the local Lorentz frame of any galaxy coincides locally with the hypersurface (see Fig. 3.1). In other words, all the local Lorentz frames of the galaxies 'mesh' together to form the hypersurface. Thus the four-velocity of a galaxy is orthogonal to the hypersurface. This series of hypersurfaces can be labelled by a parameter which may be taken as the proper time of any

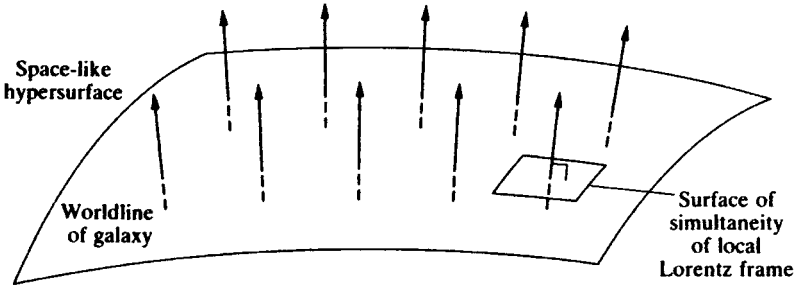


Fig. 3.1. Representation of a typical space-like hypersurface on which galaxies are assumed to lie.

galaxy, that is, time as measured by a clock stationary in the galaxy. As we shall see, this defines a universal time so that a particular time means a given space-like hypersurface on this series of hypersurfaces.

An equivalent description, known as Weyl's postulate (Weyl, 1923) is to assume that the worldlines of galaxies are a bundle or congruence of geodesics in space-time diverging from a point in the (finite or infinitely distant) past, or converging to such a point in the future, or both. These geodesics are non-intersecting, except possibly at a singular point in the past or future or both. There is one and only one such geodesic passing through each regular (that is, a point which is not a singularity) space-time point. This assumption is satisfied to a high degree of accuracy in the actual universe. The deviation from the general motion postulated here is observed to be random and small. The concept of a singular point introduced here will be elucidated in the next chapter and in Chapter 7.

We assume that the bundle of geodesics satisfying Weyl's postulate possesses a set of space-like hypersurfaces orthogonal to them. Choose a parameter t such that each of these hypersurfaces corresponds to $t =$ constant for some constant. The parameter t can be chosen to measure the proper time along a geodesic. Now introduce spatial coordinates (x^1, x^2, x^3) which are constant along any geodesic. Thus, for each galaxy the coordinates (x^1, x^2, x^3) are constant. Under these circumstances the metric can be written as follows:

$$\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 - h_{ij}\mathrm{d}x^i\mathrm{d}x^j, \quad (i, j = 1, 2, 3), \tag{3.1}$$

where the h_{ij} are functions of (t, x^1, x^2, x^3) and as usual repeated indices are to be summed over (Latin indices take values 1, 2, 3). The fact that the metric given by (3.1) incorporates the properties described above can be seen as follows. Let the worldline of a galaxy be given by $x^\mu(\tau)$, where τ is

the proper time along the galaxy. Then according to our assumptions $x^\mu(\tau)$ is given as follows:

$$(x^0 = c\tau,\ x^1 = \text{constant},\ x^2 = \text{constant},\ x^3 = \text{constant}). \tag{3.2}$$

From (3.1) and (3.2) we see that the proper time τ along the galaxy is, in fact, equal to the coordinate time t. This is because from (3.2) $\mathrm{d}x^i = 0$ along the worldline so that putting $\mathrm{d}x^i = 0$ in (3.1) yields $\mathrm{d}s = c\,\mathrm{d}\tau = c\,\mathrm{d}t$, so that $\tau = t$. Clearly a vector along the worldline given by $A^\mu = (c\,\mathrm{d}t, 0, 0, 0)$ and the vector $B^\mu = (0, \mathrm{d}x^1, \mathrm{d}x^2, \mathrm{d}x^3)$ lying in the hypersurface $t =$ constant are orthogonal, that is,

$$g_{\mu\nu}A^\mu B^\nu = 0, \tag{3.3}$$

since $g_{0i} = 0$ $(i = 1, 2, 3)$ in the metric given by (3.1). Further, the worldline given by (3.2) satisfies the geodesic equation

$$\frac{\mathrm{d}^2x^\mu}{\mathrm{d}s} + \Gamma^\mu_{\lambda\nu}\frac{\mathrm{d}x^\lambda}{\mathrm{d}s}\frac{\mathrm{d}x^\nu}{\mathrm{d}s} = 0. \tag{3.4}$$

This can be seen from the fact that, from (3.2), we have

$$\mathrm{d}x^\mu/\mathrm{d}s = (1, 0, 0, 0) \tag{3.5}$$

so that (3.4) is satisfied if $\Gamma^\mu_{00} = 0$. In fact

$$\Gamma^\mu_{00} = \tfrac{1}{2}g^{\mu\nu}(2g_{\nu 0,0} - g_{00,\nu}). \tag{3.6}$$

Using the fact that $g^{0i} = 0$ $(i = 1, 2, 3)$ which follows from (3.1), it is readily verified that Γ^μ_{00} given by (3.6) vanishes, so that (3.4) is satisfied and that the worldlines given by (3.2) are indeed geodesics.

The metric given by (3.1) does not incorporate the property that space is homogeneous and isotropic. This form of the metric can be used, with the help of a special coordinate system obtained by singling out a particular typical galaxy, to derive some general properties of the universe without the assumptions of homogeneity and isotropy (see, for example, Raychaudhuri (1955)). We shall be concerned with this general form in Chapter 7, but here we consider the form taken by (3.1) when space is homogeneous and isotropic.

The spatial separation on the same hypersurface $t =$ constant of two nearby galaxies at coordinates (x^1, x^2, x^3) and $(x^1 + \Delta x^1, x^2 + \Delta x^2, x^3 + \Delta x^3)$ is

$$\mathrm{d}\sigma^2 = h_{ij}\Delta x^i \Delta x^j. \tag{3.7}$$

Consider the triangle formed by these nearby galaxies at some particular time, and the triangle formed by these same galaxies at some later time. By the postulate of homogeneity and isotropy all points and directions on a

particular hypersurface are equivalent, so that the second triangle must be similar to the first one and further, the magnification factor must be independent of the position of the triangle in the three-space. It follows that the functions h_{ij} must involve the time coordinate t through a common factor so that ratios of small distances are the same at all times. Thus the metric has the form

$$\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 - R^2(t)\gamma_{ij}\mathrm{d}x^i\,\mathrm{d}x^j, \tag{3.8}$$

where the γ_{ij} are functions of (x^1, x^2, x^3) only. Consider the three-space given by

$$\mathrm{d}\sigma'^2 = \gamma_{ij}\mathrm{d}x^i\,\mathrm{d}x^j. \tag{3.9}$$

We assume this three-space to be homogeneous and isotropic. According to a theorem of differential geometry, this must be a space of constant curvature (see, for example, Eisenhart (1926) or Weinberg (1972)). In such a space the Riemann tensor can be constructed from the metric (and not its derivatives) and constant tensors only. The following three-dimensional fourth rank tensor constructed out of the three-dimensional metric tensor of (3.9) has the correct symmetry properties for the Riemann tensor:

$$^{(3)}R_{ijkl} = k(\gamma_{ik}\gamma_{jl} - \gamma_{il}\gamma_{jk}), \tag{3.10}$$

where k is a constant. One can verify that the three-dimensional Riemann tensor of the space given by (3.9) has the form (3.10) if the γ_{ij} are chosen to be given by the following metric (Weinberg 1972, Chapter 13):

$$\begin{aligned}\mathrm{d}\sigma'^2 &= (1+\tfrac{1}{4}kr'^2)^{-2}[(\mathrm{d}x^1)^2 + (\mathrm{d}x^2)^2 + (\mathrm{d}x^3)^2],\\ r'^2 &= (x^1)^2 + (x^2)^2 + (x^3)^2.\end{aligned} \tag{3.11}$$

The metric (3.8) can then be written as follows:

$$\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 - \frac{R^2(t)(\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2)}{[1+\tfrac{1}{4}k(x^2+y^2+z^2)]^2}, \tag{3.12}$$

where we have set $x^1 = x$, $x^2 = y$, $x^3 = z$, so that $r'^2 = x^2 + y^2 + z^2$. With $x = r'\sin\theta\cos\phi$, $y = r'\sin\theta\sin\phi$, $z = r'\cos\theta$, (3.12) reduces to the following:

$$\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 - R^2(t)\left[\frac{\mathrm{d}r'^2 + r'^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)}{(1+\tfrac{1}{4}kr'^2)^2}\right]. \tag{3.13}$$

The transformation $r = r'/(1+\tfrac{1}{4}kr'^2)$ yields the standard form of the Robertson–Walker metric, as follows:

$$\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 - R^2(t)\left[\frac{\mathrm{d}r^2}{1-kr^2} + r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)\right]. \tag{3.14}$$

The constant k in (3.14) can take the values -1, 0, $+1$, giving three different kinds of spatial metrics. We will deal with these in detail later.

We will now give a brief discussion of the manner in which the Robertson–Walker metric is derived more rigorously with the help of Killing vectors. A space is said to be homogeneous if there exists an infinitesimal isometry of the metric which can carry any point into any other point in its neighbourhood. From the discussion of Killing vectors it follows that this implies the existence of Killing vectors of the metric which at any point can take all possible values. These remarks can be illustrated by a simple example. Consider the following metric:

$$\mathrm{d}s^2 = A(t)\,\mathrm{d}t^2 - B(t)\,\mathrm{d}x^2 - C(t)\,\mathrm{d}y^2 - D(t)\,\mathrm{d}z^2, \tag{3.15}$$

where A, B, C, D are functions of the time coordinate t only, and x, y, z are the spatial coordinates. Consider two arbitrary points P and P' with spatial coordinates (a, b, c) and (a', b', c') respectively. Consider now the transformation given by

$$x' = x + a' - a, \quad y' = y + b' - b, \quad z' = z + c' - c. \tag{3.16}$$

This transformation takes the point P to the point P', because when $(x, y, z) = (a, b, c)$, we get $(x', y', z') = (a', b', c')$. On the other hand the new metric is given by

$$\mathrm{d}s^2 = A(t)\,\mathrm{d}t^2 - B(t)\,\mathrm{d}x'^2 - C(t)\mathrm{d}y'^2 - D(t)\,\mathrm{d}z'^2, \tag{3.17}$$

which has the same form in the new coordinates as (3.15) has in the old coordinates. Thus (3.16) represents an isometry of the metric, which is not just infinitesimal but a finite or a global isometry. Thus the metric (3.15) represents a homogeneous space. In terms of Killing vectors, it is easily verified that the vectors given by $\xi^\mu = (0, 1, 0, 0)$, $\eta^\mu = (0, 0, 1, 0)$ and $\zeta^\mu = (0, 0, 0, 1)$ are all Killing vectors, as are any linear combinations of these with arbitrary constant coefficients. One can thus get Killing vectors which take arbitrary spatial values, which correspond to isometries of the form (3.16).

One can similarly define isotropy in terms of isometries and Killing vectors. A space is isotropic at a point X if there exists an infinitesimal isometry which leaves the point X unchanged but takes any direction at X to any other direction, that is, takes any infinitesimal vector at X to any other one. In terms of Killing vectors, this implies the existence of Killing vectors which vanish at X but whose derivatives can take all possible values, subject to Killing's equation. The metric (3.15), although homogeneous, is not in general isotropic. A space is isotropic if it is isotropic

about every point in it. Proceeding along these lines one can derive the Robertson–Walker metric with the use of Killing vectors. We will carry out such a derivation later in this chapter.

3.2 Some geometric properties of the Robertson–Walker metric

Consider the Robertson–Walker metric (3.14) when $k = 1$. This yields the universe with positive spatial curvature whose spatial volume is finite, as we shall see. In this case it is convenient to introduce a new coordinate ψ by the relation $r = \sin\psi$, so that the metric (3.14) becomes

$$\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 - R^2(t)[\mathrm{d}\psi^2 + \sin^2\psi(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)]. \tag{3.18}$$

Some insight may be gained by embedding the spatial part of this metric in a four-dimensional Euclidean space. In general a three-dimensional Riemannian space with a positive definite metric cannot be embedded in a four-dimensional Euclidean space, but the spatial part of (3.18) can, in fact, be so embedded. Before proceeding to do this, we consider a simple example of embedding, namely, that of the space given by the two-dimensional metric

$$\mathrm{d}\sigma'^2 = a^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2). \tag{3.19}$$

This, of course, is just the surface of a two-sphere and is represented by the equation $x^2 + y^2 + z^2 = a^2$ in ordinary three-dimensional Euclidean space. This is a trivial example of the embedding of the two-surface given by (3.19). However, a metric such as (3.19) describes the intrinsic properties of the surface and does not depend on its embedding in a higher-dimensional space, although in this simple case it is natural to think in terms of the surface of an ordinary sphere in three dimensions. Turning to (3.18), we write the spatial part as follows:

$$\mathrm{d}\sigma^2 = R^2[\mathrm{d}\psi^2 + \sin^2\psi(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)], \tag{3.20}$$

where we concentrate on a particular time t and regard R as constant. Consider now a four-dimensional Euclidean space with coordinates (w, x, y, z) which are Cartesian-like in that the distance between points given by (w_1, x_1, y_1, z_1) and (w_2, x_2, y_2, z_2) is Σ_{12}, where

$$\Sigma_{12}^2 = (w_1 - w_2)^2 + (x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2. \tag{3.21}$$

Thus the metric in this space is given by

$$\mathrm{d}\Sigma^2 = \mathrm{d}w^2 + \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2. \tag{3.22}$$

Consider now a surface in this space given parametrically by

$$w = R\cos\psi, \quad x = R\sin\psi\sin\theta\cos\phi,$$
$$y = R\sin\psi\sin\theta\sin\phi, \quad z = R\sin\psi\cos\theta, \tag{3.23}$$

from which we get

$$w^2 + x^2 + y^2 + z^2 = R^2. \tag{3.24}$$

Evaluating $\mathrm{d}w$, $\mathrm{d}x$, $\mathrm{d}y$, $\mathrm{d}z$ in terms of $\mathrm{d}\psi$, $\mathrm{d}\theta$, $\mathrm{d}\phi$ from (3.23) and substituting in (3.22) we get precisely the metric given by (3.20). Just as all points and all directions starting from a point on a two-sphere in a three-dimensional Euclidean space are equivalent, so all points and directions on a three-sphere in a four-dimensional Euclidean space are equivalent. This can be seen from the fact that rotations in the four-dimensional embedding space (which can be affected by a 4×4 orthogonal matrix) can move any point and any direction on the three-sphere into any other point and direction respectively, while leaving unchanged the metric (3.22) and the equation of the three-sphere (3.24). This shows that the metric (3.20), that is, the space $t =$ constant in (3.18), is indeed homogeneous and isotropic.

Consider again a particular time t so that R can be taken as constant in (3.23) and (3.24). Consider the two-surface given by $\psi =$ constant $= \psi_0$, which is a two-sphere, as can be seen from (3.23) and (3.24), whence we get $w = R\cos\psi_0$, and

$$x^2 + y^2 + z^2 = R^2\sin^2\psi_0. \tag{3.25}$$

The surface area of this two-sphere is $4\pi R^2\sin^2\psi_0$. As ψ_0 ranges from 0 to π, one moves outwards from the 'north pole' (given by $\psi_0 = 0$) of the hypersurface through successive two-spheres of area $4\pi R^2\sin^2\psi_0$. The area increases until $\psi_0 = \pi/2$, after which it decreases until it is zero at $\psi_0 = \pi$. The distance from the 'north' to the 'south pole' is $R\pi$. This behaviour is similar to what happens on a two-sphere in a three-dimensional Euclidean space, as illustrated in Fig. 3.2. Suppose the radius of the two-sphere is R' and ψ' denotes the co-latitude. The circumference of the circle on the sphere given by $\psi' =$ constant $= \psi'_0$ is $2\pi R'\sin\psi'_0$, while the distance of this circle from the north pole O is $R'\psi'_0$. The circumference of this circle reaches a maximum at $\psi'_0 = \pi/2$, after which it decreases until it reaches zero at $\psi'_0 = \pi$, when the distance from the north pole along the surface is $R'\pi$, analogously to the previous case.

In the case of the three-space (3.24), the entire surface is swept by the coordinate range $0 \leqslant \psi \leqslant \pi$, $0 \leqslant \theta \leqslant \pi$, $0 \leqslant \phi \leqslant 2\pi$. The total volume of the three-space (3.20) is

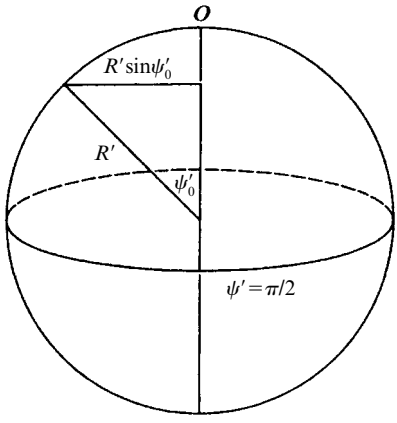


Fig. 3.2. Diagram to illustrate the analogy between the surface of a two-sphere and three-space of positive curvature.

$$\int (-{}^{(3)}g)^{1/2} \mathrm{d}^3x = \int (R\,\mathrm{d}\psi)(R\sin\psi\,\mathrm{d}\theta)(R\sin\psi\sin\theta\,\mathrm{d}\phi) = 2\pi^2 R^3, \tag{3.26}$$

which is finite. Here ${}^{(3)}g$ is the determinant of the three-dimensional metric.

In the case $k = 0$ the spatial metric is given by

$$\mathrm{d}\sigma_1^2 = R^2[\mathrm{d}\psi^2 + \psi^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)], \tag{3.27}$$

which is the ordinary three-dimensional Euclidean space. As usual, the transformation

$$x = R\psi\sin\theta\cos\phi, \quad y = R\psi\sin\theta\sin\phi, \quad z = R\psi\cos\theta, \tag{3.28}$$

gives

$$\mathrm{d}\sigma_1^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2. \tag{3.29}$$

The range of (ψ, θ, ϕ) is $0 \leqslant \psi < \infty$, $0 \leqslant \theta \leqslant \pi$, $0 \leqslant \phi \leqslant 2\pi$, and the spatial volume is infinite. This is also referred to as the universe with zero spatial curvature, as opposed to the case $k = 1$, which has positive spatial curvature.

The case $k = -1$ corresponds to the universe with negative spatial curvature. The spatial part of this metric cannot be embedded in a four-dimensional Euclidean space, but it can be embedded in a four-dimensional Minkowski space. It is, in fact, the space-like surface given by

$$x^2 + y^2 + z^2 - w^2 = -R^2, \tag{3.30}$$

in the Minkowski space with metric

$$ds^2 = dw^2 - dx^2 - dy^2 - dz^2. \tag{3.31}$$

Putting $k = -1$ and $r = \sinh\psi$ in (3.14), we get for the spatial part of this metric the following form:

$$d\sigma_2^2 = R^2[d\psi^2 + \sinh^2\psi(d\theta^2 + \sin^2\theta\, d\phi^2)]. \tag{3.32}$$

To see the embedding given by (3.30), (3.31) we transform to a Minkowski space with coordinates (w, x, y, z) given by

$$\begin{aligned} w &= R\cosh\psi, \quad x = R\sinh\psi\sin\theta\cos\phi, \\ y &= R\sinh\psi\sin\theta\sin\phi, \quad z = R\sinh\psi\cos\theta, \end{aligned} \tag{3.33}$$

which gives (3.30) on substitution for w, x, y, z. Evaluating dw, dx, dy, dz from (3.33) in terms of $d\psi, d\theta, d\phi$, and substituting in (3.31) we get the metric (3.32). In this case the surface $w = \text{constant}$ given by $\psi = \text{constant} = \psi_0$, corresponds, by substituting into (3.30) $w = R\cosh\psi_0$, to the surface of the two-sphere given by

$$x^2 + y^2 + z^2 = R^2\sinh^2\psi_0. \tag{3.34}$$

The surface of this sphere has area $4\pi R^2\sinh^2\psi_0$, which keeps on increasing indefinitely as ψ_0 increases. As is clear from the metric (3.32), the 'radius' of this sphere, that is, the distance from the 'centre' given by $\psi = 0$ to the surface given by $\psi = \psi_0$ along $\theta = $ constant and $\phi = $ constant, is $R\psi_0$. Thus the surface area is larger than that of a sphere of radius $R\psi_0$ in Euclidean space. In this case the range of the coordinates (ψ, θ, ϕ) is: $0 \leqslant \psi \leqslant \infty, 0 \leqslant \theta \leqslant \pi, 0 \leqslant \phi < 2\pi$. The spatial volume is infinite.

3.3 Some kinematic properties of the Robertson–Walker metric

We have seen that galaxies have fixed spatial coordinates, that is, they are at rest in the coordinate system defined above. Such a system is called comoving. Thus the cosmological 'fluid' is at rest in the comoving frame we have chosen. We now consider the behaviour of a free particle which is travelling with respect to this comoving frame. It is free in the sense that it is affected only by the 'background' cosmological gravitational field and no other forces. This could be a projectile shot out of a galaxy or a light wave (photon) travelling through intergalactic space. Consider the Robertson–Walker metric in the form

$$ds^2 = c^2\, dt^2 - R^2(t)\left[\frac{dr^2}{1 - kr^2} + r^2(d\theta^2 + \sin^2\theta\, d\phi^2)\right]. \tag{3.35}$$

We write $(x^0, x^1, x^2, x^3) = (ct, r, \theta, \phi)$, so that

$$g_{00} = 1, \quad g_{11} = -R^2(t)/(1-kr^2),$$
$$g_{22} = -R^2(t)r^2, \quad g_{33} = -R^2(t)r^2\sin^2\theta, \tag{3.36}$$

the rest of the metric components being zero. Consider a geodesic passing through a typical point P. Without loss of generality we can take the spatial origin of the coordinate system, that is, $r = 0$, to be at the point P. The path of the particle is given by the geodesic equation

$$\frac{\mathrm{d}u^\mu}{\mathrm{d}\lambda} + \Gamma^\mu_{\alpha\beta}u^\alpha u^\beta = 0, \tag{3.37}$$

where $u^\mu = \mathrm{d}x^\mu/\mathrm{d}\lambda$, $x^\mu(\lambda)$, being the coordinates of a space-time point on the worldline of the moving particle as a function of the path parameter λ. If the particle is massive, λ can be taken as the proper time s of the particle, and if it is a photon, λ is an affine parameter.

Multiply (3.37) by $g_{\sigma\mu}$ and use (2.4), (2.7) to get

$$g_{\sigma\mu}(\mathrm{d}u^\mu/\mathrm{d}\lambda) + \tfrac{1}{2}(g_{\sigma\alpha,\beta} + g_{\sigma\beta,\alpha} - g_{\alpha\beta,\sigma})u^\alpha u^\beta = 0. \tag{3.38}$$

We also have

$$\frac{\mathrm{d}u_\sigma}{\mathrm{d}\lambda} = \frac{\mathrm{d}}{\mathrm{d}\lambda}(g_{\sigma\mu}u^\mu) = g_{\sigma\mu}\frac{\mathrm{d}u^\mu}{\mathrm{d}\lambda} + g_{\sigma\mu,\rho}u^\rho u^\mu. \tag{3.39}$$

In (3.38) $g_{\sigma\alpha,\beta}u^\alpha u^\beta = g_{\sigma\beta,\alpha}u^\alpha u^\beta$, so that if we eliminate from this equation the term $g_{\sigma\mu}\,\mathrm{d}u^\mu/\mathrm{d}\lambda$ with the use of (3.39), we arrive at the following equation

$$\mathrm{d}u_\sigma/\mathrm{d}\lambda - \tfrac{1}{2}g_{\alpha\beta,\sigma}u^\alpha u^\beta = 0. \tag{3.40}$$

Equation (3.40) tells us that if the metric components are independent of a particular coordinate x^σ, then the covariant component u_σ is constant along the geodesic. Consider the component $\sigma = 3$, so that we are referring to $x^3 = \phi$. Since the metric components (3.36) are independent of ϕ, we have $\mathrm{d}u_3/\mathrm{d}\lambda = 0$, so that u_3 is constant along the geodesic. But

$$u_3 = g_{33}u^3 = -R^2(t)r^2(\sin^2\theta)u^3, \tag{3.41}$$

so that $u_3 = 0$ at the point P were $r = 0$. Thus $u_3 = 0$ along the geodesic and so $u^3 = \mathrm{d}\phi/\mathrm{d}\lambda = 0$ as well, so ϕ is constant along the geodesic.

Consider (3.40) for $\sigma = 2$:

$$\mathrm{d}u_2/\mathrm{d}\lambda - \tfrac{1}{2}g_{\alpha\beta,2}u^\alpha u^\beta = 0. \tag{3.42}$$

The only component of $g_{\alpha\beta}$ which depends on $x^2 = \theta$ is g_{33}, but the contribution of the corresponding term to (3.42) vanishes since $u^3 = 0$. Thus $\mathrm{d}u_2/\mathrm{d}\lambda = 0$, so u_2 is constant along the geodesic. Again

$$u_2 = g_{22}u^2 = -R^2(t)r^2u^2, \tag{3.43}$$

which vanishes at P ($r = 0$), and so u_2 is zero along the geodesic, as is u^2, so that θ is also constant along the geodesic.

To proceed further we concentrate on the case $k = 0$ in (3.35) and (3.36). We leave it as an exercise for the reader to extend the following analysis to the cases $k = +1, -1$. In these two cases it is helpful to transform the coordinate r to ψ given by $r = \sin\psi$, $r = \sinh\psi$ respectively, as in (3.20), (3.32). We return to (3.40) with $\sigma = 1$:

$$\mathrm{d}u_1/\mathrm{d}\lambda - \tfrac{1}{2}g_{\alpha\beta,\,1}u^\alpha u^\beta = 0. \tag{3.44}$$

We have $u^2 = u^3 = 0$, while g_{00} and g_{11} are independent of r (recall that $k = 0$). Thus $\mathrm{d}u_1/\mathrm{d}\lambda = 0$ so that u_1 is constant along the geodesic:

$$u_1 = g_{11}u^1 = -R^2(t)\frac{\mathrm{d}r}{\mathrm{d}s} = \text{constant}, \tag{3.45}$$

where we have taken the parameter λ to be the proper time s. In the metric (3.35) we can set $\mathrm{d}\theta = \mathrm{d}\phi = 0$ (since θ and ϕ are constant along the geodesic) to get

$$\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 - R^2(t)\,\mathrm{d}r^2 = c^2\,\mathrm{d}t^2 - \mathrm{d}l^2 = \mathrm{d}t^2(c^2 - v^2), \tag{3.46}$$

where $\mathrm{d}l$ is the element of spatial distance and $v = \mathrm{d}l/\mathrm{d}t$ is the velocity of the particle in the comoving frame, assuming it to be a massive particle of mass m. The momentum of the particle is given as follows:

$$q = m\,(\mathrm{d}l/\mathrm{d}s)c = mv/(1 - v^2/c^2)^{1/2}. \tag{3.47}$$

Combining (3.45), (3.46), (3.47) we get

$$qR(t) = \text{constant along the geodesic}. \tag{3.48}$$

The above analysis can also be applied to the case of a photon, in which case, since the energy q_0 and the momentum q of the photon are related by $q_0 = cq$, we have

$$q_0R(t) = \text{constant along the geodesic}. \tag{3.49}$$

Since the energy of the photon is proportional to its frequency ν, we get

$$\nu R(t) = \text{constant along the geodesic}. \tag{3.50}$$

Consider a photon emitted at time t_1 with frequency ν_1 which is observed at the point P at time t_0 with frequency ν_0. From (3.50) we get

$$\nu_0 R(t_0) = \nu_1 R(t_1). \tag{3.51}$$

This can be written as

$$1 + z = R(t_0)/R(t_1), \tag{3.52}$$

where $z = (\lambda_0 - \lambda_1)/\lambda_1$ is the fractional change in the wavelength; λ_0, λ_1 being the wavelengths corresponding to the frequencies ν_0, ν_1 (with $\nu_0\lambda_0 = \nu_1\lambda_1 = c$). The number z is always observed to be positive, at least for distant galaxies, indicating a shift in the visible spectrum towards red, so that z is referred to as the 'red-shift'. We will come back to (3.52) later, but now we discuss another derivation of this relation.

The light ray follows a path given by $\mathrm{d}s = 0$, which, with the use of (3.46), yields the following relation

$$\int_{t_1}^{t_0} c(dt/R(t)) = \int_0^{r_1} \mathrm{d}r = r_1, \tag{3.53}$$

assuming the emitting galaxy to be at $r = r_1$. If the next wave train leaves the galaxy at $t_1 + \delta t_1$ and arrives at $t_0 + \delta t_0$, (3.53) implies

$$\int_{t_1}^{t_0} c\,\mathrm{d}t/R(t) = \int_{t_1+\delta t_1}^{t_0+\delta t_0} c\,\mathrm{d}t/R(t). \tag{3.54}$$

Assuming δt_0, δt_1 to be small compared to t_0, t_1, (3.54) can be approximated as follows

$$\delta t_1/R(t_1) = \delta t_0/R(t_0). \tag{3.55}$$

Since the frequency is inversely proportional to the time interval in which the wave train is emitted, we get (3.51) again.

Without any further consideration the function $R(t)$ which occurs in the Robertson–Walker metric can be *any* function of the time t. From (3.52) we see, since z is observed not to be zero, that the function $R(t)$ is not just a constant. To determine this function we must resort to dynamics, which are provided by Einstein's equations. Before considering these, in the next chapter, we discuss some further properties of the Robertson–Walker metric which are independent of what form the function $R(t)$ takes. These properties may be referred to as kinematic properties.

As indicated in Chapter 1, the first evidence of a systematic red-shift in the spectra of light coming from distant galaxies was found by Hubble. He analysed the data on frequency shifts obtained earlier by Slipher and

others and found a linear relationship between the red-shift z and the distance l. He interpreted the red-shift as being due to the recessional velocity of the galaxies. The approximate argument, which is valid if the values of the red-shifts are not high, goes as follows. Let δt_1 of the earlier discussion following (3.53) represent the time interval during which successive wave crests leave the source at $r = r_1$, and let δt_0 be the interval during which these wave crests are received by the observer. If the source is moving away from the observer with velocity v, during the time the two consecutive wave crests are emitted the source moves a distance $v\delta t_1$. Because of this movement, the time interval in which the crests reach the observer is increased by an amount $v\delta t_1/c$. Thus we have

$$\delta t_0 = \delta t_1 + v\delta t_1/c. \tag{3.56}$$

The wavelengths of the emitted and observed light are given as follows:

$$\lambda_1 = c\delta t_1, \quad \lambda_0 = c\delta t_0. \tag{3.57}$$

From (3.56) and (3.57) it follows that

$$\lambda_0/\lambda_1 = \delta t_0/\delta t_1 = 1 + v/c = 1 + z. \tag{3.58}$$

Thus $z = v/c$. This is true if the velocity is small compared to the speed of light. From (3.52) and (3.58) we get

$$v = cz = c(t_0 - t_1)\dot{R}(t_1)/R(t_1), \tag{3.59}$$

where we have assumed $t_0 - t_1$ to be small and expanded $R(t)$ about $t = t_1$, with $\dot{R}(t) \equiv \mathrm{d}R(t)/\mathrm{d}t$. Again if $t_0 - t_1$ is small the t_1 in the arguments of R and $\dot{R}$ in (3.59) can be replaced by t_0. With the use of similar approximations, we derive the following relations between the coordinate distance r_1 and the distance l of the galaxy:

$$r_1 = c(t_0 - t_1)/R(t_0), \tag{3.60}$$

$$l = r_1 R(t_0) = c(t_0 - t_1). \tag{3.61}$$

With the use of (3.59), (3.60) and (3.61) we finally get Hubble's law, as follows:

$$v = cz = H_0 l, \quad H_0 = \dot{R}(t_0)/R(t_0). \tag{3.62}$$

There are many uncertainties in the exact determination of Hubble's constant, H_0, some of which we shall discuss later in the book. One of the best values available for some years was that of Sandage and Tammann (1975), as follows (other measurements will be mentioned later):

$$H_0 = (50.3 \pm 4.3)\ \mathrm{km\ s^{-1}\ Mpc^{-1}}. \tag{3.63}$$

Here Mpc stands for megaparsec, which is approximately 3.26 million light years.

As mentioned already, the formula (3.62) holds only when the red-shift is small. We should expect departures from this linear Hubble's law if the red-shift is not small. To this end, we expand $R(t)$ in a Taylor series about the *present* epoch t_0, as follows:

$$\begin{aligned} R(t) &= R[t_0 - (t_0 - t)] \\ &= R(t_0) - (t_0 - t)\dot{R}(t_0) + \tfrac{1}{2}(t_0 - t)^2 \ddot{R}(t_0) - \cdots \\ &= R(t_0)[1 - (t_0 - t)H_0 - \tfrac{1}{2}(t_0 - t)^2 q_0 H_0^2 - \cdots], \end{aligned} \tag{3.64}$$

with

$$q_0 = -\ddot{R}(t_0)R(t_0)/\dot{R}^2(t_0). \tag{3.65}$$

With the use of (3.53) with a minor adjustment of sign we get

$$\begin{aligned} r = \int_t^{t_0} c\,\mathrm{d}t/R(t) &= \int_t^{t_0} c\,\mathrm{d}t/\{R(t_0)[1 - (t_0 - t)H_0 - \cdots]\} \\ &= cR^{-1}(t_0)[(t_0 - t) + \tfrac{1}{2}(t_0 - t)^2 H_0 + \cdots]. \end{aligned} \tag{3.66}$$

Here r is the coordinate radius of the galaxy under consideration. The first term in the last expression in (3.66) gives (3.60). With the use of the first part of (3.61), namely, $l = rR(t_0)$, we can invert (3.66) to obtain $t_0 - t$ in terms of l as follows

$$t_0 - t = l/c - \tfrac{1}{2}H_0 l^2/c^2. \tag{3.67}$$

From (3.52) and (3.64) we can find z up to second order in $t_0 - t$ as follows:

$$\begin{aligned} z &= [1 - (t_0 - t)H_0 - \tfrac{1}{2}(t_0 - \mathrm{t})^2 q_0 H_0^2 - \cdots]^{-1} - 1 \\ &= (t_0 - t)H_0 + (t_0 - \mathrm{t})^2(\tfrac{1}{2}q_0 + 1)H_0^2 + \cdots. \end{aligned} \tag{3.68}$$

We now substitute for $t_0 - t$ from (3.67) into (3.68) to obtain a relation for the red-shift z in terms of the distance l.

$$z = H_0 l/c + \tfrac{1}{2}(1 + q_0)H_0^2 l^2/c^2 + O(H_0^3 l^3). \tag{3.69}$$

Thus from the observed red-shifts it is possible to determine the parameters H_0 and q_0 if an independent estimate can be obtained for the distance. The parameter q_0 is referred to as the deceleration parameter, as it indicates by how much the expansion of the universe is slowing down. If the expansion is speeding up, for which there appears to be some recent evidence, then q_0 will be negative.

3.4 The Einstein equations for the Robertson–Walker metric

In this section we derive the Einstein equations given by (2.22) for the Robertson–Walker metric, in which the matter is in the form of a perfect fluid of mass-energy density ε and pressure p, so that the energy–momentum tensor is given by (2.23), with $u^\mu = (1, 0, 0, 0)$, as we are in comoving coordinates.

The metric components and Christoffel symbols which are non-zero are given as follows (recall that $(x^0, x^1, x^2, x^3) = (ct, r, \theta, \phi)$):

$$g_{00} = 1, \quad g_{11} = -R^2/(1-kr^2), \quad g_{22} = -r^2R^2,$$
$$g_{33} = \sin^2\theta R^2, \tag{3.70}$$
$$g^{00} = 1, \quad g^{11} = -(1-kr^2)/R^2, \quad g^{22} = -(rR)^{-2},$$
$$g^{33} = -(r\sin\theta R)^{-2}. \tag{3.71}$$

We put the Christoffel symbols $\Gamma^\mu_{\nu\lambda}$ in four groups according to the values 0, 1, 2, 3 of the index μ, as follows:

$$\Gamma^0_{11} = c^{-1}R\dot{R}/(1-kr^2), \quad \Gamma^0_{22} = c^{-1}r^2R\dot{R},$$
$$\Gamma^0_{33} = c^{-1}r^2\sin^2\theta R\dot{R}, \tag{3.72a}$$
$$\Gamma^1_{01} = c^{-1}\dot{R}/R, \quad \Gamma^1_{11} = kr/(1-kr^2), \quad \Gamma^1_{22} = -r(1-kr^2),$$
$$\Gamma^1_{33} = -r(1-kr^2)\sin^2\theta, \tag{3.72b}$$
$$\Gamma^2_{02} = c^{-1}\dot{R}/R, \quad \Gamma^2_{12} = 1/r, \quad \Gamma^2_{33} = -\sin\theta\cos\theta, \tag{3.72c}$$
$$\Gamma^3_{03} = c^{-1}\dot{R}/R, \quad \Gamma^3_{13} = 1/r, \quad \Gamma^3_{23} = \cot\theta. \tag{3.72d}$$

We next substitute the Christoffel symbols into (2.17) or (2.18) to get the following non-zero components of the Ricci tensor $R_{\mu\nu}$ (note that r is dimensionless while $R(t)$ has the dimension of length).

$$R_{00} = -3\ddot{R}/R, \tag{3.73a}$$
$$R_{11} = (R\ddot{R} + 2\dot{R}^2 + 2c^2k)/(1-kr^2), \tag{3.73b}$$
$$R_{22} = r^2(R\ddot{R} + 2\dot{R}^2 + 2c^2k), \tag{3.73c}$$
$$R_{33} = r^2\sin^2\theta(R\ddot{R} + 2\dot{R}^2 + 2c^2k). \tag{3.73d}$$

It is unfortunate that the same letter is normally used for the scale factor $R(t)$ as for the Ricci scalar (see (2.20)), but it should be clear from the context which is meant. The Ricci scalar can be evaluated with the use of (3.73a)–(3.73d) as follows:

$$g^{\mu\nu}R_{\mu\nu} = -6(R\ddot{R} + \dot{R}^2 + c^2k)/R^2. \tag{3.74}$$

We are now in a position to write down the Einstein equations (2.22), noting that the covariant components of the four-velocity are the same as the contravariant ones: $u_\mu = (1, 0, 0, 0)$, so that the non-zero components of $T_{\mu\nu}$ are:

$$T_{00} = \varepsilon, \quad T_{11} = pR^2/(1 - kr^2), \quad T_{22} = pr^2R^2,$$
$$T_{33} = pr^2(\sin^2\theta)R^2. \tag{3.75}$$

The 00- and 11-components of (2.22) can be written as follows:

$$3(\dot{R}^2 + c^2k) = 8\pi G\varepsilon R^2/c^2, \tag{3.76a}$$
$$2R\ddot{R} + \dot{R}^2 + kc^2 = -8\pi GpR^2/c^2. \tag{3.76b}$$

The 00-component of (2.22) has been multiplied by R^2 to get (3.76a), while the 11-component has been multiplied by $kr^2 - 1$ to get (3.76b). The 22- and 33-components of (2.22) yield equations which are equivalent to (3.76b).

A useful consequence of (3.76a) and (3.76b) can be obtained by considering the equation of conservation of mass-energy given by (2.25). A generalization of (2.6a) implies that (2.25) can be written as follows:

$$T^{\mu\nu}{}_{,\nu} + \Gamma^\mu_{\nu\sigma}T^{\sigma\nu} + \Gamma^\nu_{\nu\sigma}T^{\mu\sigma} = 0. \tag{3.77}$$

With the non-zero contravariant components $T^{\mu\nu}$ given as follows:

$$T^{00} = \varepsilon, \quad T^{11} = p(1 - kr^2)/R^2, \quad T^{22} = p/(rR)^2,$$
$$T^{33} = p/[r(\sin\theta)R]^2, \tag{3.78}$$

and with the use of the Christoffel symbols (3.72a)–(3.72d), (3.77) can be written as follows:

$$\dot{\varepsilon} + 3(p + \varepsilon)\dot{R}/R = 0, \tag{3.79}$$

which comes from the $\mu = 0$ component of (3.77), the other components being satisfied identically. Equation (3.79) is, of course, a consequence of (3.76a) and (3.76b) and can be derived from these by first evaluating $\dot{\varepsilon}$ from (3.76a) and using (3.76b) to eliminate $\ddot{R}$.

In the next section we shall attempt to provide the essential framework in which a rigorous derivation of the Robertson–Walker metric can be carried out (Weinberg, 1972, Chapter 13). We will mention the construction briefly; the details are given by Weinberg.

3.5 Rigorous derivation of the Robertson–Walker metric

Consider Killing's equation (2.38) and a given point X with coordinates X^μ which is situated in a neighbourhood with coordinates x^μ. By a 'neighbourhood' we mean a region of space-time in which all points can be represented by the same coordinate system; this is also referred to as a 'coordinate patch'. As mentioned earlier, several coordinate patches may be required for a global description, that is, to describe the whole of space-time. The Killing condition is such that it enables one to calculate the function $\xi_\mu(x)$ in the whole neighbourhood from the values of $\xi_\mu(x)$ and the derivatives $\xi_{\mu;\nu}(x)$ at the point X, i.e., from the quantities $\xi_\mu(X)$, $\xi_{\mu;\nu}(X)$. In these arguments x, X represent x^μ, X^μ respectively. When we say 'derivatives' here, it makes no difference if we mean ordinary or covariant derivatives, because the latter is expressible in terms of the former and components of the vector itself, at the point X (see (2.6b)). The fact that $\xi_\mu(x)$ is determined thus can be seen as follows.

We write down (2.38) here for convenience:

$$\xi_{\rho;\sigma} + \xi_{\sigma;\rho} = 0. \tag{3.80}$$

Using ξ_μ instead of A_μ in (2.12) we get

$$\xi_{\mu;\nu;\lambda} - \xi_{\mu;\lambda;\nu} = \xi_\sigma R^\sigma{}_{\mu\nu\lambda}. \tag{3.81}$$

The following relation is obtained by raising the index σ in (2.14c):

$$R^\sigma{}_{\mu\nu\lambda} + R^\sigma{}_{\lambda\mu\nu} + R^\sigma{}_{\nu\lambda\mu} = 0. \tag{3.82}$$

Equation (3.81) implies, taking cyclic permutations of $(\nu\lambda\mu)$, the following two equations:

$$\xi_{\nu;\lambda;\mu} - \xi_{\nu;\mu;\lambda} = \xi_\sigma R^\sigma{}_{\nu\lambda\mu}, \tag{3.83a}$$

$$\xi_{\lambda;\mu;\nu} - \xi_{\lambda;\nu;\mu} = \xi_\sigma R^\sigma{}_{\lambda\mu\nu}. \tag{3.83b}$$

Adding (3.81), (3.83a,b) and using (3.82), we get

$$\xi_{\mu;\nu;\lambda} - \xi_{\mu;\lambda;\nu} + \xi_{\nu;\lambda;\mu} - \xi_{\nu;\mu;\lambda} + \xi_{\lambda;\mu;\nu} - \xi_{\lambda;\nu;\mu} = 0. \tag{3.84}$$

Taking covariant derivatives of Killing's equation (3.80) with suitable combinations of indices, we get the following three equations, if ξ_μ satisfies (3.80):

$$\xi_{\mu;\nu;\lambda} + \xi_{\nu;\mu;\lambda} = 0, \tag{3.85a}$$

$$\xi_{\mu;\lambda;\nu} + \xi_{\lambda;\mu;\nu} = 0, \tag{3.85b}$$

$$\xi_{\nu;\lambda;\mu} + \xi_{\lambda;\nu;\mu} = 0. \tag{3.85c}$$

With the use of these equations, (3.84) reduces to the following equation:

$$\xi_{\mu;\nu;\lambda} - \xi_{\mu;\lambda;\nu} - \xi_{\lambda;\nu;\mu} = 0. \tag{3.86}$$

Using (3.81) again we get

$$\xi_{\lambda;\nu;\mu} = \xi_\sigma R^\sigma{}_{\mu\nu\lambda}. \tag{3.87}$$

With the use of (2.6b) and the following corresponding relation for a second rank covariant tensor $A_{\lambda\nu}$:

$$A_{\lambda\nu;\mu} = A_{\lambda\nu,\mu} - \Gamma^\sigma_{\lambda\mu} A_{\sigma\nu} - \Gamma^\sigma_{\nu\mu} A_{\lambda\sigma}, \tag{3.88}$$

it is readily verified that (3.87) can be written as follows:

$$\begin{aligned}\xi_{\lambda,\nu,\mu} = \xi_\sigma R^\sigma{}_{\mu\nu\lambda} &+ (\Gamma^\sigma_{\lambda\nu,\mu} - \Gamma^\rho_{\lambda\mu}\Gamma^\sigma_{\rho\nu} - \Gamma^\rho_{\nu\mu}\Gamma^\sigma_{\lambda\rho})\xi_\sigma \\ &+ (\Gamma^\sigma_{\lambda\nu}\xi_{\sigma,\mu} + \Gamma^\sigma_{\lambda\mu}\xi_{\sigma,\nu} + \Gamma^\sigma_{\mu\nu}\xi_{\lambda,\sigma}).\end{aligned} \tag{3.89}$$

This relation shows that, in the neighbourhood of X, the second derivatives of ξ_λ can be expressed in terms of the ξ_λ and its first derivatives. Consider now any function $f(x)$ of the coordinates x^μ in the neighbourhood of X. By a Taylor series, $f(x)$ can be expanded as follows around X^μ:

$$\begin{aligned} f(x) &= f(X + x - X) \\ &= f(X) + (x^\mu - X^\mu)(\partial f(x)/\partial x^\mu) \\ &\quad + \tfrac{1}{2}(x^\mu - X^\mu)(x^\nu - X^\nu)(\partial^2 f/\partial x^\mu \partial x^\nu) + \cdots. \end{aligned} \tag{3.90}$$

In a similar manner the vector $\xi_\mu(x)$ can be expanded around X^μ:

$$\begin{aligned} \xi_\mu(x) &= \xi_\mu(X + x - X) \\ &= \xi_\mu(X) + (x^\sigma - X^\sigma)(\partial \xi_\mu/\partial x^\sigma) \\ &\quad + \tfrac{1}{2}(x^\sigma - X^\sigma)(x^\rho - X^\rho)(\partial^2 \xi_\mu/\partial x^\sigma x^\rho) + \cdots. \end{aligned} \tag{3.91}$$

Note that the derivatives are evaluated at $x^\mu = X^\mu$. It is clear from (3.89) that by repeated use of this equation and its derivatives all higher derivatives of ξ_μ can be expressed in terms of ξ_μ and its first derivatives. Thus all the second and higher derivatives of ξ_μ occurring in (3.91) (evaluated at X) can be expressed in terms of $\xi_\mu(X)$ and $\xi_{\mu;\nu}(X)$. Therefore, in the neighbourhood of X, $\xi_\mu(x)$ turns out to be a linear combination of $\xi_\mu(X)$ and $\xi_{\mu;\nu}(X)$, with coefficients which are functions of x^μ, as well as X^μ. This relation can be expressed as follows:

$$\xi^{(n)}_\mu(x) = A^\lambda_\mu(x;X)\xi^{(n)}_\lambda(X) + B^{\lambda\nu}_\mu(x;X)\xi^{(n)}_{\lambda;\nu}(X), \tag{3.92}$$

where, for convenience, we have reverted to $\xi_{\lambda;\nu}$ in place of $\xi_{\lambda,\nu}$, since the two derivatives are equivalent for the present purpose; changing $\xi_{\lambda;\nu}$ to $\xi_{\lambda,\nu}$ simply changes the coefficient $A^{\lambda}_{\mu}(x;X)$ in a manner which is readily determined. Note also that the coefficients $A^{\lambda}_{\mu}(x;X)$ and $B^{\lambda\nu}_{\mu}(x;X)$ are both functions of x^{μ} and X^{μ}; these can be determined explicitly in terms of $R^{\sigma}{}_{\mu\nu\lambda}$, the Γs and their derivatives evaluated at X; $\xi_{\mu}(x)$, $\xi_{\mu,\nu}(x)$ can be expressed as power series in $(x^{\mu}-X^{\mu})$ with the use of (3.89) and its derivatives, and of (3.91). The superscript n in $\xi^{(n)}_{\mu}$, etc., in (3.92) indicates the different possible linearly independent solutions of Killing's equation that can exist at X. Thus every Killing vector $\xi_{\mu}(x)$ is determined through (3.92) in a neighbourhood in terms of the values of $\xi_{\mu}(x)$ and $\xi_{\mu;\nu}(x)$ at any given point X of the neighbourhood. If the different Killing vectors $\xi^{(n)}_{\mu}$, $n=1,2,\ldots$ satisfy an equation

$$\sum_{n} c_n \xi^{(n)}_{\mu}(x) = 0, \tag{3.93}$$

with constant coefficients c_n, then they are linearly dependent; otherwise they are linearly independent. Note that the above equations are valid in any number N of dimensions. From (3.92) it therefore follows that in N dimensions there can be at most $\frac{1}{2}N(N+1)$ linearly independent Killing vectors in any neighbourhood. This can be seen as follows. From (3.92) we see that $\xi^{(n)}_{\mu}(x)$ is linearly dependent on $\xi^{(n)}_{\mu}(X)$ and $\xi^{(n)}_{\mu;\nu}(X)$ for any n and x. Now in N dimensions, for any n there are $\frac{1}{2}N(N-1)$ quantities $\xi^{(n)}_{\mu;\nu}(X)$, and together with the N quantities $\xi^{(n)}_{\mu}(X)$, these give $N+\frac{1}{2}N(N-1)=\frac{1}{2}N(N+1)$ independent quantities. For different values of n these can be regarded as vectors in a $\frac{1}{2}N(N+1)$ dimensional space. In such a space there can be at most $\frac{1}{2}N(N+1)$ linearly independent vectors. Thus for any fixed x, X, (3.92) can yield at most $\frac{1}{2}N(N+1)$ linearly independent vectors $\xi^{(n)}_{\mu}(x)$. This argument may be a little more transparent if it is stated as follows. Let us fix x and X, and let the indices $(\lambda\nu)$ in $B^{\lambda\nu}_{\mu}$ be denoted by Λ which takes values $1,2,\ldots,\frac{1}{2}N(N-1)$, and let $\xi^{(n)}_{\mu}(X)$, $\xi^{(n)}_{\lambda;\nu}(X)$ be denoted respectively by $\hat{\xi}^{(n)}_{\mu}$, $\hat{\xi}'^{(n)}_{\Lambda}$, the 'hat' on ξ or ξ' denoting that these quantities are evaluated at X. Then (3.92) can be written as follows:

$$\xi^{(n)}_{\mu}(x) = A^{\lambda}_{\mu}\hat{\xi}^{(n)}_{\lambda} + B^{\Lambda}_{\mu}\hat{\xi}'^{(n)}_{\Lambda}, \tag{3.94}$$

where λ, as usual, takes values $1,2,\ldots,N$ while Λ takes values $1,2,\ldots,\frac{1}{2}N(N-1)$. Since x, X are fixed, the A^{λ}_{μ} and B^{Λ}_{μ} can be regarded as constants. Let there be M such vectors, for $n=1,2,\ldots,M$. Thus for fixed x, X, for any given n, $\xi^{(n)}_{\mu}(x)$ is a linear combination of $\frac{1}{2}N(N+1)$ quantities, namely $\hat{\xi}^{(n)}_{\lambda}$, $\lambda=1,2,\ldots,N$, $\hat{\xi}'^{(n)}_{\Lambda}$, $\Lambda=1,2,\ldots,\frac{1}{2}N(N-1)$, which quantities together can be regarded as a vector in a $\frac{1}{2}N(N+1)$ dimensional space, namely, the vector

$$(\hat{\xi}^{(n)}_1, \ldots, \hat{\xi}^{(n)}_N, \hat{\xi}^{\prime(n)}_1, \ldots, \hat{\xi}^{\prime(n)}_{\frac{1}{2}N(N-1)}). \tag{3.95}$$

Clearly these are $\frac{1}{2}N(N+1)$ linearly independent vectors; for example, they can be taken as the $\frac{1}{2}N(N+1)$ linearly independent vectors (1,0, . . .,0), (0,1,0, . . .,0), . . ., (0, . . .,0,1,0), (0, . . .,0,1), each of these having $\frac{1}{2}N(N+1)-1$ zeros as components. Hence at any point x there can be at most $\frac{1}{2}N(N+1)$ linearly independent vectors, since for $M>\frac{1}{2}N(N+1)$ we must have, for any M vectors of the form (3.95) with $n=1,2, \ldots, M$, the following relations holding for some constants $c_1, c_2, \ldots, c_M$:

$$\sum_{n=1}^{M} c_n \hat{\xi}_1{}^{(n)} = 0, \quad \sum_{n=1}^{M} c_n \hat{\xi}_2{}^{(n)} = 0, \ldots, \quad \sum_{n=1}^{M} c_n \hat{\xi}^{(n)}_N = 0,$$

$$\sum_{n=1}^{M} c_n \hat{\xi}^{\prime(n)}_1 = 0, \ldots, \quad \sum_{n=1}^{M} c_n \hat{\xi}^{\prime(n)}_{\frac{1}{2}N(N-1)} = 0. \tag{3.96}$$

These equations imply that the $\xi^{(n)}_\mu(x)$ must satisfy the following equations (for $M>\frac{1}{2}N(N+1)$):

$$\sum_{n=1}^{M} c_n \xi^{(n)}_\mu(x) = 0. \tag{3.97}$$

That is, the resulting $\xi^{(n)}_\mu(x)$ (for $n=1,2, \ldots, M$) are linearly dependent, for $M>\frac{1}{2}N(N+1)$.

As mentioned earlier, a space is homogeneous if there are infinitesimal isometries (2.36) that can carry any point X to any other point in its neighbourhood. This is equivalent to saying that Killing's equation (3.80) has solutions, for the given metric and the given point, which can take all possible values.

For greater generality we continue to work in N dimensions. Let there be N Killing vectors

$$\xi^{(\mu)}_\lambda(x;X), \quad \mu=1,2, \ldots, N, \tag{3.98a}$$

with

$$\xi^{(\mu)}_\lambda(X;X), \quad \delta^\mu_\lambda, \tag{3.98b}$$

where δ^μ_λ is the Kronecker delta in N dimensions. The $\xi^\mu_\lambda(x;X)$ thus defined are linearly independent, for if we had

$$\sum_{\mu=1}^{M} c_\mu \xi^{(\mu)}_\lambda(x;X) = 0, \tag{3.99a}$$

then, putting $x^\mu = X^\mu$ in (3.99a), we get, using (3.98b),

$$\sum_{\mu=1}^{M} c_\mu \xi^{(\mu)}_\lambda(X;X) = \sum_{\mu=1}^{M} c_\mu \delta^\mu_\lambda = c_\lambda = 0. \tag{3.99b}$$

A metric space is isotropic about any point X of it if there are infinitesimal isometries (2.36) which leave X unchanged: $\xi^{\lambda}(X)=0$, while the derivatives $\xi_{\lambda;\nu}(X)$ can take all possible values. If the space has N dimensions, it is possible to choose $\frac{1}{2}N(N-1)$ Killing vectors $\xi_{\lambda}^{(\mu\nu)}(x;X)$ satisfying the following relations:

$$\xi_{\lambda}^{(\mu\nu)}(x;X)=-\xi_{\lambda}^{(\nu\mu)}(x;X), \tag{3.100a}$$

$$\xi_{\lambda}^{(\mu\nu)}(X;X)=0, \tag{3.100b}$$

$$\begin{aligned}\xi_{\lambda;\rho}^{(\mu\nu)}(X;X)&=[(\partial/\partial x^{\rho})\xi_{\lambda}{}^{(\mu\nu)}(x;X)]_{x=X}\\&=\delta_{\lambda}^{\mu}\delta_{\rho}^{\nu}-\delta_{\rho}^{\mu}\delta_{\lambda}^{\nu}.\end{aligned} \tag{3.100c}$$

(Equation (3.100b) implies that ordinary and covariant derivatives coincide at $x=X$). The $\frac{1}{2}N(N-1)$ Killing vectors defined as above are linearly independent, for let

$$\sum_{\mu,\nu} c_{\mu\nu}\xi_{\lambda}{}^{(\mu\nu)}(x;X)=0, \tag{3.101a}$$

with $c_{\mu\nu}=-c_{\nu\mu}$. Since this is meant to be an identity, its derivative with respect to x^{σ} should be valid; taking this derivative and setting $x=X$ and using the condition (3.100c) we get

$$c_{\lambda\sigma}-c_{\sigma\lambda}=2c_{\lambda\sigma}=0, \tag{3.101b}$$

which proves linear independence.

Spaces which are isotropic about *every* point are of particular interest in cosmology. It is sufficient to consider isotropy about two infinitesimally near points X^{μ}, $X^{\mu}+\mathrm{d}X^{\mu}$, so that there exist Killing vectors

$$\xi_{\lambda}^{(\mu\nu)}(x;X), \quad \xi_{\lambda}^{(\mu\nu)}(x;X+\mathrm{d}X), \tag{3.102}$$

that vanish at $x=X$, $x=X+\mathrm{d}X$ respectively (see (3.100b)), and such that the derivatives take all possible values. Since (3.102) are both Killing vectors, any linear combination is also a Killing vector, and so

$$\operatorname*{Lim}_{\mathrm{d}X^{\rho}\to 0}\{[\xi_{\lambda}^{(\mu\nu)}(x+\hat{\mathrm{d}}X;X)-\xi_{\lambda}^{(\mu\nu)}(x;X)]/\mathrm{d}X^{\rho}\}=\partial\xi_{\lambda}^{(\mu\nu)}(x;X)/\partial X^{\rho}, \tag{3.103}$$

is also a Killing vector, where $\hat{\mathrm{d}}X$ means $\hat{\mathrm{d}}X^{\mu}=0$ if $\mu\neq\rho$, $\hat{\mathrm{d}}X^{\rho}\neq 0$. We evaluate this quantity at $x=X$ as follows. First note that $\xi_{\lambda}^{(\mu\nu)}(X;X)$ vanishes identically (see (3.100b)). Therefore the derivative of this function of X (or this set of functions, to be more precise) with respect to X^{ρ} also vanishes identically:

$$\partial\xi_{\lambda}^{(\mu\nu)}(X;X)/\partial X^{\rho}=0. \tag{3.104}$$

But

$$\begin{aligned}\partial\xi_\lambda^{(\mu\nu)}(X;X)/\partial X^\rho &= \lim_{\mathrm{d}X^\rho\to 0}\{(X+\hat{\mathrm{d}}X;X+\hat{\mathrm{d}}X)-\xi_\lambda^{(\mu\nu)}(X;X)]/\mathrm{d}X^\rho\}\\ &= \lim_{\mathrm{d}X^\rho\to 0}\{[\xi_\lambda^{(\mu\nu)}(X+\hat{\mathrm{d}}X;X+\hat{\mathrm{d}}X)-\xi_\lambda^{(\mu\nu)}(X+\hat{\mathrm{d}}X;X)\\ &\qquad +\xi_\lambda^{(\mu\nu)}(X+\hat{\mathrm{d}}X;X)-\xi_\lambda^{(\mu\nu)}(X;X)]/\mathrm{d}X^\rho\}\\ &= \{\partial\xi_\lambda^{(\mu\nu)}(x;X)/\partial X^\rho\}_{x=X}+\{\partial\xi_\lambda^{(\mu\nu)}(x;X)/\partial x^\rho\}_{x=X},\end{aligned} \tag{3.105}$$

where $\hat{\mathrm{d}}X$ has the same meaning as above. From (3.104) and (3.105) we get

$$\begin{aligned}\{\partial\xi_\lambda^{(\mu\nu)}(x;X)/\partial X^\rho\}_{x=X} &= -\{\partial\xi_\lambda^{(\mu\nu)}(x;X)/\partial x^\rho\}_{x=X}\\ &= -\delta_\lambda^\mu\delta_\rho^\nu+\delta_\rho^\mu\delta_\lambda^\nu,\end{aligned} \tag{3.106}$$

where in the last step we have used (3.100c) and (3.103). Thus, from the fact that there exist Killing vectors (3.102) that vanish at infinitesimally nearby points X and $X+\mathrm{d}X$ and whose derivatives can take any values, we can construct a Killing vector $\xi_\lambda(x)$, which can take any arbitrary value a_λ, as follows. We define $\xi_\lambda(x)$ by the following equation:

$$\xi_\lambda(x)=(a_\nu/(N-1))(\partial\xi_\lambda^{(\rho\nu)}(x;X)/\partial X^\rho), \tag{3.107}$$

where $\xi_\lambda^{(\rho\nu)}(x;X)$ has been defined as in (3.100a,b,c). We have already seen that $\partial\xi_\lambda^{(\rho\nu)}(x;X)/\partial X^\rho$ is a Killing vector, and from (3.106) we see that

$$\begin{aligned}\xi_\lambda(X) &= (a_\nu/(N-1))\{\partial\xi_\lambda^{(\rho\nu)}(x;X)/\partial X^\rho\}_{x=X}\\ &= (a_\nu/(N-1))(-\delta_\lambda^\rho\delta_\rho^\nu+\delta_\rho^\rho\delta_\lambda^\nu)\\ &= (a_\nu/(N-1))(-\delta_\lambda^\nu+N\delta_\lambda^\nu)=a_\lambda.\end{aligned} \tag{3.108}$$

Thus a space which is isotropic about every point of it is also homogeneous, because the latter condition amounts to the existence of Killing vectors that can take any arbitrary values, which (3.108) implies.

As a 'comprehension exercise', the reader may wish to carry out the above analysis explicitly for the case $N=4$, in which case the number of independent Killing vectors is 10, arising essentially out of the four $\xi_\mu(X)$ and the six independent $\xi_{\mu;\nu}(X)$.

In this section we have followed closely the account of this topic given by Weinberg (1972, Chapter 13), but the treatment here is more detailed and explicit in places.

We have seen that there can be at most $\frac{1}{2}N(N+1)$ independent Killing vectors in N dimensions. A metric which admits the maximum number is referred to by Weinberg as *maximally symmetric*. Such a space is necessarily homogeneous and isotropic about every point. Maximally symmetric spaces are uniquely determined by the 'curvature constant' K and the signature of the metric (number of positive and negative terms in the diagonal form). In various cases of physical interest, the whole of space (or space-time) is not maximally symmetric, but it may be possible to decompose it into such subspaces. A spherically symmetric three-dimensional space, for example, can be decomposed into a series of subspaces, each being maximally symmetric in two dimensions (see (3.19)). In cosmology we have one example of a maximally symmetric space-time: the de Sitter or the steady state universe given by (9.13). More importantly, we have space-times in which each 'plane' of constant time is maximally symmetric. We refer to Weinberg (1972, Chapter 13) for the detailed construction of such metrics, including the Robertson–Walker metric, from this point of view. Weinberg's discussion of these matters is very instructive for the serious student of cosmology; the considerations of this section may provide a useful background. The papers by Robertson (1935, 1936) and by Walker (1936) are outstanding landmarks in the theoretical development of modern cosmology.

4

The Friedmann models

4.1 Introduction

In Section 3.4 we derived the Einstein equations for the Robertson–Walker metric with the energy–momentum tensor as that of a perfect fluid in which the matter is at rest in the local frame. While the Robertson–Walker metric incorporates the symmetry properties and the kinematics of space-time, the Einstein equations provide the dynamics, that is, the manner in which the matter, and the space-time in turn, are affected by the forces present in the universe.

We rewrite (3.76a) and (3.76b) as follows. First we eliminate $\dot{R}^2$ from (3.76b) to get the following equation:

$$\ddot{R} = -(4\pi G/3)(\varepsilon + 3p)R/c^2. \tag{4.1}$$

Next we write (3.76a) for the three different values of k: -1, 0, 1.

$$\dot{R}^2 = c^2 + (8\pi G/3)\varepsilon R^2/c^2, \tag{4.2a}$$

$$\dot{R}^2 = (8\pi G/3)\varepsilon R^2/c^2, \tag{4.2b}$$

$$\dot{R}^2 = -c^2 + (8\pi G/3)\varepsilon R^2/c^2. \tag{4.2c}$$

For any one of the three values of k, we have two equations for the three unknown functions R, ε, p. We need one more equation, which is provided by the equation of state, $p = p(\varepsilon)$, in which the pressure is given as a function of the mass-energy density. With the equation of state given, the problem is determinate and the three functions R, ε, p can be worked out completely. Models of the universe which are determined in this way are referred to as Friedmann models, after the Russian mathematician A. A. Friedmann (1888–1925) who was the first to study these models.

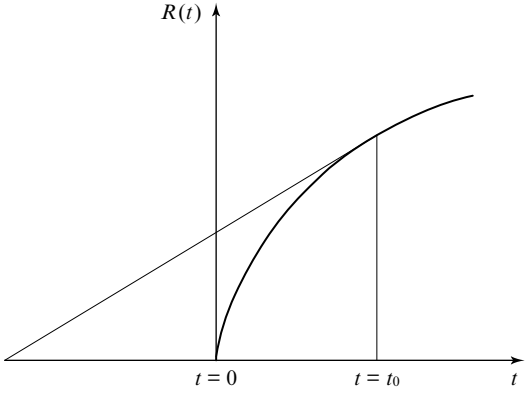


Fig. 4.1. Diagram to illustrate (4.4), that is, the result that the age of the universe is less than the Hubble time.

Some information can be obtained about the function $R(t)$ without solving the equations explicitly, if one makes a few reasonable assumptions about the pressure and density. For example, if we assume that $\varepsilon + 3p$ remains positive, then from (4.1) we see that the 'acceleration' $\ddot{R}/R$ is negative. Let the present time be denoted by $t = t_0$. Now $R(t_0) > 0$ (by definition) and $\dot{R}(t_0)/R(t_0) > 0$ (because we see red-shifts, not blue-shifts – see (3.59)); it follows that the curve $R(t)$ must be concave downwards (towards the t-axis – see Fig. 4.1). It is also clear from the figure that the curve $R(t)$ must reach the t-axis at a time which is closer to the present time than the time at which the tangent to the point $(t_0, R(t_0))$ reaches the t-axis. We refer to the time at which $R(t)$ reaches the t-axis as $t = 0$. Thus at a finite time in the past, namely $t = 0$, we have

$$R(0) = 0. \tag{4.3}$$

The point $t = 0$ can reasonably be called the beginning of the universe. Clearly, the point at which the tangent meets the t-axis is the point at which $R(t)$ would have been zero if the expansion had been uniform, that is, if $\dot{R}$ was constant and $\ddot{R} = 0$. The time elapsed from that point till the present time is $R(t_0)/\dot{R}(t_0) = H_0^{-1}$ (see the discussion on page 8). Thus, since, in fact, $\ddot{R}$ is negative for $0 < t < t_0$, it follows that the age of the universe must be less than the Hubble time:

$$t_0 < H_0^{-1}. \tag{4.4}$$

Adding $\dot{p}$ to both sides of (3.79) and multiplying the resulting equation by R^3, we get the following equation:

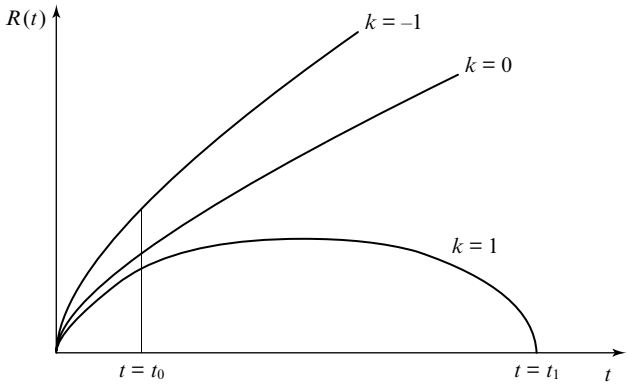


Fig. 4.2. The behaviour of the curve $R(t)$ for the three values -1, 0, $+1$ of k. The time $t=t_0$ is the present time and $t=t_1$ the time at which $R(t)$ reaches zero again for $k=+1$.

$$\dot{p}R^3=\frac{\mathrm{d}}{\mathrm{d}t}[R^3(\varepsilon+p)]. \tag{4.5}$$

We multiply (4.5) by $\dot{R}^{-1}$ and transform the derivative with respect to t to a derivative with respect to R, to arrive at the following equation:

$$\frac{\mathrm{d}}{\mathrm{d}R}(\varepsilon R^3)=-3pR^2. \tag{4.6}$$

From this equation we see that as long as the pressure p remains positive, the density ε must decrease with increasing R at least as fast as R^{-3}. This is because if the pressure is zero in (4.6), the density varies exactly as R^{-3}, and with a negative right hand side (for positive pressure), the density must decrease faster than R^{-3}. Thus as R tends to infinity, the quantity εR^2 vanishes at least as fast as R^{-1}. We see that in the cases $k=-1$ and $k=0$, given respectively by (4.2a) and (4.2b), $\dot{R}^2$ remains positive definite so that $R(t)$ keeps on increasing. From (4.2a) we clearly get the result

$$R(t)\to ct \quad \text{as} \quad t\to\infty;\, k=-1. \tag{4.7}$$

For $k=0$ also, $R(t)$ goes on increasing, but more slowly than t. In the case $k=+1$, given by (4.2c), $\dot{R}^2$ becomes zero when εR^2 reaches the value $3c^4/8\pi G$. Since $\ddot{R}$ is negative definite, the curve $R(t)$ must continue to be concave towards the t-axis, so that $R(t)$ begins to decrease, and must reach $R(t)=0$ at some finite time in the future (the time $t=t_1$ in Fig. 4.2). The three cases $k=-1$, 0, $+1$ are illustrated in Fig. 4.2.

In (3.63) we have mentioned approximately 50 km s^{-1} Mpc^{-1} as a possible value of Hubble's constant. To find the corresponding Hubble time,

we merely have to determine the time taken to traverse 1 Mpc at a speed of 50 km s^{-1}. Since there are about 3×10^{19} km in a megaparsec and about 3×10^{7} s in a year, we readily see that the value of 50 for H_0 in the above units corrresponds to a Hubble time of about 20 billion years. Similarly, a value of 100 for H_0 gives a Hubble time of approximately 10 billion years.

Recalling that $H_0=\dot{R}(t_0)/R(t_0)$ (see (3.62)), the three relations (4.2a)–(4.2c) (that is, (3.76a)) can be written as follows:

$$3c^2H_0^2/(8\pi G)=-3kc^4/(8\pi GR_0^2)+\varepsilon_0, \tag{4.8}$$

where $R_0=R(t_0)$. Let us denote the left hand side of (4.8) by ε_c, and call it the critical density, ε_0 being the present value of the density. From (4.8) we see that if ε_0 is less than ε_c, then k is negative, whereas if ε_0 is greater than ε_c, then k is positive. From the above discussion (see Fig. 4.2) we see that the universe will expand forever if the present density is below the critical density, and it will stop expanding and collapse to zero $R(t)$ at some time in the future if the present density is above the critical density. With $G=6.67\times10^{-8}$ dyne cm^2 g^{-2}, the value of the critical density can be written as follows:

$$\varepsilon_c/c^2\equiv3H_0^2/(8\pi G)=4.9\times10^{-30}(H_0/50\text{ km s}^{-1}\text{ Mpc}^{-1})^2\text{ g cm}^{-3}. \tag{4.9}$$

Thus if H_0 has the value 50 in the usual units, the critical density is approximately five times 10^{-30} g cm^{-3}, or, since the proton mass is about 1.67×10^{-24} g, about three hydrogen atoms in every thousand litres of space, as mentioned in Chapter 1. From (4.8) and (4.9) one gets $\Omega_0=(8\pi G\varepsilon_0/3c^2H_0^2)$, the present value of the density parameter Ω introduced in Chapter 1.

Recalling the definition of the deceleration parameter q_0 (see (3.65)), and denoting by a subscript zero all quantities evaluated at the present epoch $t=t_0$, from (4.1) we get

$$\varepsilon_0+3p_0=-3\ddot{R}_0c^2/(4\pi GR_0)=(3/4\pi G)q_0H_0^2c^2. \tag{4.10}$$

We next eliminate ε_0 between (4.8) and (4.10) to get the following expression for p_0:

$$p_0=-(8\pi G)^{-1}[kc^2/R_0^2+H_0^2(1-2q_0)]c^2. \tag{4.11}$$

Observationally it is found that the present universe is dominated by non-relativistic matter, that is,

$$p_0\ll\varepsilon_0, \tag{4.12}$$

so that if p_0 is negligible we get from (4.11)

$$c^2k/R_0^2 = (2q_0 - 1)H_0^2. \tag{4.13}$$

From (4.8), (4.9) and (4.13) we then get the following simple relation between the ratio of the present density to the critical density and the deceleration parameter:

$$\varepsilon_0/\varepsilon_c = 2q_0. \tag{4.14}$$

We see from (4.14) or directly from (4.13) that the universe is open if q_0 is less than $\frac{1}{2}$ and closed if it is greater than $\frac{1}{2}$. If the present density is exactly equal to the critical density or if q_0 is exactly equal to $\frac{1}{2}$ (together with the assumption of zero pressure in the latter case), we have $k = 0$ and the universe is open (see Fig. 4.2).

4.2 Exact solution for zero pressure

As we have noted, observationally the pressure seems to be negligible compared to the mass-energy density. We shall discuss this further later, but for the present we set $p = 0$, because this yields an exact solution for all time, and, although it may not be accurate, especially for the early epoch of the universe, it provides a useful model. In this case (4.6) can be integrated at once to yield the following equation:

$$\varepsilon/\varepsilon_0 = (R_0/R)^3. \tag{4.15}$$

We eliminate ε_0 and k/R_0^2 with the use of (4.10) (recalling that $p_0 = 0$) and (4.13), and use (4.15) to write (3.76a) as follows:

$$(\dot{R}/R_0)^2 = H_0^2(1 - 2q_0 + 2q_0R_0/R). \tag{4.16}$$

The solution of this equation can be expressed as an integral, giving t in terms of R, as follows:

$$t = H_0^{-1} \int_0^R (1 - 2q_0 + 2q_0R_0/R')^{-1/2}\, \mathrm{d}R'/R_0, \tag{4.17}$$

with $t = 0$ being the value of t for which $R(t) = 0$. In particular, the present age of the universe is obtained by taking R_0 as the upper limit in the integral in (4.17). This age can be expressed in terms of H_0 and q_0, both of which are observational parameters, by changing the variable of integration to $w = R'/R_0$, as follows:

$$t_0 = H_0^{-1} \int_0^1 (1 - 2q_0 + 2q_0/w)^{-1/2}\, \mathrm{d}w. \tag{4.18}$$

This relation holds for all three values of k, but with the assumption of zero pressure. It is clear that for any positive q_0, the present age t_0 given by (4.18) must satisfy the inequality (4.4). We now consider explicitly three different cases, denoted by (i), (ii), and (iii) below.

(i) $k=+1$, $\varepsilon_0>\varepsilon_c$.

From (4.14) we see that this case corresponds to $q_0>\frac{1}{2}$. In this case the integral in (4.17) can be integrated by the following substitution:

$$1-\cos\theta=(q_0R_0)^{-1}(2q_0-1)R', \tag{4.19}$$

the resulting equation being given by the following:

$$H_0t=q_0(2q_0-1)^{-3/2}(\theta-\sin\theta). \tag{4.20}$$

After the integration the R' in (4.19) can be replaced by $R(t)$. Equations (4.19) and (4.20) then imply that the curve $R(t)$ is a cycloid. From the left hand side of (4.19) it is clear that $R(t)$ increases from zero at $\theta=0$ to its maximum value at $\theta=\pi$, and then decreases steadily until it reaches zero again at $\theta=2\pi$. The maximum value of $R(t)$ occurs at the time T_m given by

$$T_m=\pi q_0H_0^{-1}(2q_0-1)^{-3/2},\ R(T_m)=2q_0(2q_0-1)^{-1}R_0. \tag{4.21}$$

When $R(t)$ returns to zero again, $t=2T_m$. The present value of θ, θ_0, is given by setting R' equal to R_0 in (4.19), so that

$$\cos\theta_0=q_0^{-1}-1. \tag{4.22}$$

Substituting this into (4.20) we get the present age of the universe as

$$t_0=H_0^{-1}q_0(2q_0-1)^{-3/2}[\cos^{-1}(q_0^{-1}-1)-q_0^{-1}(2q_0-1)^{1/2}]. \tag{4.23}$$

If, for example, $q_0=\frac{2}{3}$, so that $\theta_0=\pi/3$, and if H_0 is 50 in the units used earlier, so that H_0^{-1} is about 20 billion years, from (4.23) we readily see that t_0 is then approximately 12.3 billion years. In this case T_m is about 218 billion years so that the whole life cycle of the universe is about 436 billion years.

(ii) $k=0$, $\varepsilon_0=\varepsilon_c$.

From (4.14) we see that this case corrresponds to $q_0=\frac{1}{2}$. The integral (4.17) is readily evaluated to yield

$$R(t)/R_0=(3H_0t/2)^{2/3}. \tag{4.24}$$

The age of the universe is given by

$$t_0=\tfrac{2}{3}H_0^{-1}, \tag{4.25}$$

so that for the value 50 for H_0, the age is approximately 13.3 billion years. This case is known as the *Einstein–de Sitter model*.

(iii) $k=-1$, $\varepsilon_0<\varepsilon_c$.

From (4.14) it follows that this is the case $q_0<\frac{1}{2}$. The analysis of case (i) above can be taken over if we consider θ to be imaginary and set $\theta=iu$. Equation (4.20) then becomes

$$H_0t=q_0(1-2q_0)^{-3/2}(\sinh u-u), \tag{4.26}$$

where u is given by

$$\cosh u-1=(q_0R_0)^{-1}(1-2q_0)R(t). \tag{4.27}$$

As in case (ii), $R(t)$ increases without limit. For large t and u these two variables are related approximately as

$$t=H_0^{-1}q_0(1-2q_0)^{-3/2}\exp(u), \tag{4.28}$$

so that as t tends to infinity

$$R(t)/R_0\to\tfrac{1}{2}q_0(1-2q_0)^{-1}\exp(u)\to\tfrac{1}{2}(1-2q_0)^{1/2}H_0t. \tag{4.29}$$

The present value of u, u_0 is obtained by setting R equal to R_0 in (4.27) and is given as follows:

$$\cosh u_0=q_0^{-1}-1. \tag{4.30}$$

Substituting this into (3.26), we get the age of the universe as follows:

$$t_0=H_0^{-1}[(1-2q_0)^{-1}-q_0(1-2q_0)^{-3/2}\cosh^{-1}(q_0^{-1}-1)]. \tag{4.31}$$

The mass density of the visible matter, that is, the matter that is contained within the galaxies, is between a tenth and a fifth of the critical density for any reasonable value of Hubble's constant. In this case, if one takes as an example the value 0.014 for q_0, we get u_0 to be approximately 5 and then t_0 is nearly $0.96H_0^{-1}$, that is, nearly equal to the Hubble time.

The deceleration parameter q_0 provides a measure of the slowing down of the expansion of the universe. This dimensionless parameter can, of course, be defined for any time t, and in that case it could be called the deceleration function $q(t)$ (see (3.65)):

$$q(t)=-\ddot{R}(t)R(t)/\dot{R}^2(t). \tag{4.32}$$

Convenient expressions can be found for $q(t)$ in terms of the parameters θ and u introduced above in the cases $k=1$ and $k=-1$ respectively.

Consider the case $k=1$. In this case, with the use of (4.1) (with $p=0$) and (4.2c) we get

$$q(t)=(4\pi G/3)\varepsilon R^2/[-c^4+(8\pi G/3)\varepsilon R^2]. \tag{4.33}$$

Evaluating (4.33) at $t=t_0$ we get q_0 in terms of ε_0 and R_0, whence we get

$$(2q_0-1)/q_0R_0=3c^4/4\pi G\varepsilon_0R_0^3. \tag{4.34}$$

Equations (4.19) and (4.34) imply

$$(4\pi G/3c^4)\varepsilon_0R_0^3R^{-1}=(1-\cos\theta)^{-1}. \tag{4.35}$$

Eliminating ε from (4.33) with the use of (4.15) and using (4.35), we get

$$q(t)=(1+\cos\theta)^{-1}. \tag{4.36}$$

Thus as θ varies from 0 to 2π during one cycle, $q(t)$ rises from $\frac{1}{2}$ to infinity and then drops to $\frac{1}{2}$ again. An analysis similar to the one above yields for the case $k=-1$ the following expression for q:

$$q(t)=(1+\cosh u)^{-1}. \tag{4.37}$$

Thus as u varies from 0 to infinity, q decreases steadily from $\frac{1}{2}$ to zero. In the case $k=0$, we find with the use of (4.1) (with $p=0$) and (4.2b), that $q(t)$ remains constant, at the value $q=\frac{1}{2}$. There is some recent observational evidence that the deceleration parameter q_0 may be negative, that is, the universe may be accelerating. We will discuss this later.

4.3 Solution for pure radiation

When the cosmological fluid is dominated by radiation, as was presumably the case in the early universe, the equation of state can be taken as

$$p=\tfrac{1}{3}\varepsilon. \tag{4.38}$$

In this case (4.1) reduces to the following equation:

$$\ddot{R}c^2=-(8\pi G/3)\varepsilon R. \tag{4.39}$$

Equation (4.6) can now be integrated to give the relation:

$$\varepsilon/\varepsilon_0=(R_0/R)^4. \tag{4.40}$$

The equation corresponding to (4.13) can be written in this case as:

$$c^2k/R_0^2=(q_0-1)H_0^2, \tag{4.41}$$

while that corresponding to (4.16) is as follows:

$$(\dot{R}/R_0)^2=H_0^2(1-q_0+q_0R_0^2/R^2). \tag{4.42}$$

Equation (4.42) can be expressed as an integral as:

$$t = H_0^{-1} \int_0^{R/R_0} (1 - q_0 + q_0/x^2)^{-1/2}\, \mathrm{d}x, \tag{4.43}$$

with $t = 0$ being the value of t for which $R(t) = 0$. Explicit solutions can be obtained as before, but they are not of much physical interest as the present universe is far from radiation dominated. The behaviour near $t = 0$ is interesting and is considered below. One point worth noting is that in the case $k = 0$ the deceleration function $q(t)$ is constant at $q = 1$, as can be readily verified with the use of (4.2b) and (4.39).

4.4 Behaviour near $t = 0$

It is of considerable interest to determine the behaviour of the function $R(t)$ near the beginning of the universe, that is, near $t = 0$. This behaviour will be used later when we study the early universe. Consider first the zero pressure case. In this case ε varies as R^{-3} so that εR^2 varies as R^{-1}. Thus in all three cases (4.2a)–(4.2c) near $t = 0$ the following relation holds:

$$\dot{R}^2 = 2u\varepsilon_0 R_0^3 R^{-1}, \qquad u \equiv 4\pi G/3c^2. \tag{4.44}$$

In the case $k = 0$ this equation holds exactly (for zero pressure). Equation (4.44) can be integrated readily to give the following behaviour for R:

$$R(t) = (\tfrac{3}{2})^{2/3}(2u\varepsilon_0)^{1/3} R_0 t^{2/3}, \tag{4.45}$$

so that R varies as $t^{2/3}$ near $t = 0$, for zero pressure and all three values of k.

Consider next the pure radiation cases given by $p = \frac{1}{3}\varepsilon$. In this case ε varies as R^{-4} (see (4.40)), so that εR^2 varies as R^{-2}. Thus in this case too the first terms in (4.2a) and (4.2c) can be ignored near $t = 0$, and all three equations can be written as follows (with the use of (4.40)):

$$\dot{R}^2 = 2u\varepsilon_0 R_0^4 R^{-2}. \tag{4.46}$$

Again in the case $k = 0$ this equation holds exactly. This equation can be integrated to yield the following behaviour for R:

$$R(t) = (8u\varepsilon_0)^{1/4} R_0 t^{1/2}. \tag{4.47}$$

4.5 Exact solution connecting radiation and matter eras

More general equations of state than the cases of zero pressure and pure radiation mentioned above have been considered by Chernin (1965, 1968),

McIntosh (1968) and Landsberg and Park (1975). In this section we give an exact solution for an equation of state which is such that for small values of R it approximates to that of pure radiation, that is, $p=\frac{1}{3}\varepsilon$, while for large values of R the ratio between the pressure and density behaves like R^{-2}, that is, the pressure becomes negligible. We make the ansatz that the mass-energy density is given as a function of R as follows:

$$\varepsilon = AR^{-4}(R^2+b)^{1/2}, \tag{4.48}$$

where A, b are positive constants. We see that (4.48) implies that for small R, the function ε behaves like R^{-4}, while for large R it behaves like R^{-3}. These are indeed the cases of pure radiation and zero pressure, given respectively by (4.40) and (4.15). We note further that when $b=0$, (4.48) reduces to the zero pressure case given by (4.15).

We combine Equations (4.2a)–(4.2c) into the following one:

$$\dot{R}^2 = -kc^2 + 2u\varepsilon R^2, \tag{4.49}$$

(with u given by (4.44)) and substitute for ε from (4.48) to get the following equation:

$$\dot{R}^2 = -kc^2 + 2uAR^{-2}(R^2+b)^{1/2}. \tag{4.50}$$

This equation can be expressed as the following integral:

$$t = \int_0^R [-kR^2c^2 + 2uA(R^2+b)^{1/2}]^{-1/2} R\,\mathrm{d}R. \tag{4.51}$$

This integral can be simplified by the substitution:

$$x = (R^2+b)^{1/2}, \tag{4.52}$$

which transforms (4.51) as follows:

$$t = \int_{b^{1/2}}^{(R^2+b)^{1/2}} (-kx^2c^2 + 2uAx + c^2kb)^{-1/2} x\,\mathrm{d}x. \tag{4.53}$$

Consider the three cases $k=1, 0, -1$ separately.

(i) Case $k=1$.

In this case (4.53) can be integrated to yield the following parametric relation between R and t:

$$(R^2+b)^{1/2} = \frac{uA}{c^2} + \frac{\upsilon}{c}\sin\theta, \qquad \upsilon = \left(bc^2 + \frac{u^2A^2}{c^2}\right)^{1/2}, \tag{4.54a}$$

$$tc^3 = uA(\theta-\theta_0) - c\upsilon(\cos\theta - \cos\theta_0), \tag{4.54b}$$

where θ_0 is the value of θ for which R vanishes, that is,

$$b^{1/2}c^2 = uA + cv\sin\theta_0. \tag{4.54c}$$

(ii) Case $k = 0$.

In this case R and t are related as follows:

$$R^2 = -b + (wt + b^{3/4})^{4/3}, \quad w \equiv \tfrac{3}{2}(2uA)^{1/2}. \tag{4.55}$$

(iii) Case $k = -1$.

In this case (4.53) can be integrated to give the following parametric relation between R and t:

$$(R^2 + b)^{1/2} = -\frac{uA}{c^2} + \frac{v}{c}\cosh\psi, \tag{4.56a}$$

$$tc^3 = -uA(\psi - \psi_0) + cv(\sinh\psi - \sinh\psi_0), \tag{4.56b}$$

where ψ_0 is the value of ψ for which R vanishes, that is,

$$b^{1/2}c^2 = -uA + cv\cosh\psi_0. \tag{4.56c}$$

To find the pressure, we first take the derivative of (4.50) with respect to t and cancel a factor $\dot{R}$ to get the following expression for $\ddot{R}$:

$$\ddot{R} = -uAR^{-3}(R^2 + 2b)(R^2 + b)^{-1/2}. \tag{4.57}$$

From (4.1) we get p as follows:

$$3p = -(uR)^{-1}\ddot{R} - \varepsilon, \tag{4.58}$$

so that, with the use of (4.57), we arrive at the following expression for p:

$$p = (bA/3)R^{-4}(R^2 + b)^{-1/2}. \tag{4.59}$$

The equation of state is given parametrically by (4.48) and (4.59). The condition that as R tends to zero the relation between p and ε tends to $p = \frac{1}{3}\varepsilon$ is automatically satisfied by ε and p given by (4.48) and (4.59) respectively.

We get the following value for the ratio of the pressure and the mass-energy density:

$$p/\varepsilon = (b/3)(R^2 + b)^{-1}. \tag{4.60}$$

Thus near $R = 0$ this ratio is $\frac{1}{3}$ while as R tends to large values the ratio behaves as R^{-2}, that is, the pressure becomes negligible compared to the mass-energy density, as is indicated by observations.

In the case of $k=1$, we have $R=0$ at $t=0$ (for $\theta=\theta_0$), and the maximum value of R occurs at $\theta=\pi/2$, at the value of t given by

$$tc^3 = Tc^3 = uA(\pi/2-\theta_0) + cv\cos\theta_0, \tag{4.61}$$

the corresponding value of R being given by the following expression:

$$R = \left(\frac{2uA}{c^3}\right)^{1/2}\left(v+\frac{uA}{c}\right)^{1/2}. \tag{4.62}$$

After the maximum, R decreases steadily to zero in the manner of a cycloid considered earlier. This case can be considered as a generalized cycloid, the whole cycle lasting for a period of $2T$, the final value of θ being $\pi-\theta_0$. The behaviour of $R(t)$ is thus very similar to the case of pure radiation or zero pressure for $k=1$.

In all three cases $R(t)$ behaves as $t^{1/2}$ for small t, which is consistent with (4.47). In the case $k=0$ it is readily seen that for large t, $R(t)$ behaves like t. In the case $k=-1$, large values of R and t occur for large values of the parameter ψ, and for such values both Rc and tc^2 behave like $v\,e^{\psi}$, so that $R(t)$ tends to infinity like ct, in the manner of the zero pressure case with $k=-1$.

It is of some interest to note that the deceleration function, defined by (4.32), is given by the following expression for this solution:

$$q(t) = uA(R^2+2b)/\{(R^2+b)^{1/2}[-kR^2c^2+2uA(R^2+b)^{1/2}]\}. \tag{4.63}$$

The deceleration function takes the following simple form for the case $k=0$:

$$q(t) = \tfrac{1}{2}(R^2+2b)/(R^2+b). \tag{4.64}$$

This function tends to unity as R tends to zero, which is consistent with the fact that for the case of pure radiation and $k=0$, the deceleration function remains constant at the value of $q=1$ (see the end of Section 4.3). As R tends to infinity, $q(t)$ tends to $\frac{1}{2}$, consistent with the zero pressure, $k=0$ case (see the end of Section 4.2).

4.6 The red-shift versus distance relation

In Section 3.3 we considered the relation between the red-shift and distance for small values of r, $t-t_0$, l, etc. (see (3.64) (3.66), (3.67) and (3.69)). In this section we want to extend that analysis to arbitrary values of the red-shift, etc., with the use of the exact solution for zero pressure. Let a light ray emitted at $t=t_1$ from the position $r=r_1$ radially be received at the

position $r = 0$ at time $t = t_0$. Denoting by R_1 the value of R at t_1, the red-shift z is given as follows (see (3.52)):

$$1 + z = R_0/R_1. \tag{4.65}$$

We consider the analogue of (3.53) for $k \neq 0$ to get the following equation:

$$\int_0^{r_1} (1 - kr^2)^{-1/2}\, \mathrm{d}r = c \int_{t_1}^{t_0} \frac{\mathrm{d}t}{R(t)} = c \int_{R_1}^{R_0} \frac{\mathrm{d}R}{R\dot{R}}. \tag{4.66}$$

We now substitute for $\dot{R}$ from the exact solution for zero pressure given by (4.16), and transform to the integration variable $x = R/R_0$, to get

$$\int_0^{r_1} (1 - kr^2)^{-1/2}\, \mathrm{d}r = c(R_0 H_0)^{-1} \int_{(1+z)^{-1}}^{1} (1 - 2q_0 + 2q_0/x)^{-1/2} x^{-1}\, \mathrm{d}x. \tag{4.67}$$

It can be shown that for all three values of k, the expression for r_1 is the same, as follows:

$$r_1 = c\{zq_0 + (q_0 - 1)[-1 + (2q_0 z + 1)^{1/2}]\}/[H_0 R_0 q_0^2 (1 + z)]. \tag{4.68}$$

For large values of the red-shift z it is convenient to define a *luminosity distance*, measured by comparison of *apparent luminosity* and *absolute luminosity*, which are respectively the radiation received by an observer per unit area per unit time from the source, and the radiation emitted by the source per unit solid angle per unit time. The luminosity distance, d_L, is given as follows (see, for example, Weinberg (1972, p. 421)):

$$d_L = r_1 R_0^2 / R_1. \tag{4.69}$$

With the use of (4.65) and (4.68), this can be written as follows:

$$d_L = R_0 r_1 (1 + z) = c(H_0 q_0^2)^{-1}\{zq_0 + (q_0 - 1)[-1 + (2q_0 z + 1)^{1/2}]\}. \tag{4.70}$$

For small values of z we get

$$d_L = cH_0^{-1}[z + \tfrac{1}{2}(1 - q_0)z^2]. \tag{4.71}$$

This equation is independent of models and can be derived using kinematics only, like (3.69).

4.7 Particle and event horizons

In Section 4.2 we obtained exact and explicit solutions for $R(t)$ for zero pressure, that is, in the matter-dominated era. This solution can be used to illustrate certain limitations of our vision of the universe first pointed out by Rindler (1956). Here we follow closely the discussion of this question given by Weinberg (1972, p. 489). Consider an observer situated at $r=0$. Let another observer situated at $r=r_1$ emit a light signal at time t_1. Suppose this light signal reaches the first observer at time t. Assuming light to be the fastest of any signals, the only other signals emitted at time t_1 that the first observer receives by time t are from radial coordinates $r<r_1$. Extending (3.53) to the two non-zero values of k, we see that r_1 is determined as follows:

$$\int_0^{r_1} \mathrm{d}r/(1-kr^2)^{1/2} = c\int_{t_1}^{t} \mathrm{d}t'/R(t'). \tag{4.72}$$

If the t' integral in (4.72) diverges as t_1 tends to zero, then r_1 can be made as large as we please by taking t_1 to be sufficiently small. Thus in this case in principle it is possible to receive signals emitted at sufficiently early times from any comoving particle, such as a typical galaxy. If, however, the t' integral converges as t_1 tends to zero, then r_1 can never exceed a certain value for a given t. In this case our vision of the universe is limited by what Rindler has called a *particle horizon*. It is possible to receive signals at time t from comoving particles that are within the radial coordinate r_h, which is a function of t, given as follows:

$$\int_0^{r_h} \mathrm{d}r/(1-kr^2)^{1/2} = c\int_0^{t} \mathrm{d}t'/R(t'). \tag{4.73}$$

The proper distance d_h of this horizon is

$$d_\mathrm{h}(t) = R(t)\int_0^{r_h} \mathrm{d}r/(1-kr^2)^{1/2} = cR(t)\int_0^{t} \mathrm{d}t'/R(t'). \tag{4.74}$$

From (3.76a) we see that if the mass-energy density ε varies as $R^{-2-\delta}$ for some positive δ, as R goes to zero, the k on the left hand side of this equation can be neglected and it is readily seen that $R(t)$ behaves as $t^{2/(2+\delta)}$. In this case the t' integral in (4.73) converges as t_1 goes to zero and a particle horizon is present. This is the case in the solution for zero pressure considered in Section 4.2. If the largest contribution to the t' integral comes from the matter-dominated era, we can use (4.17) to express d_h as follows:

$$d_{\mathrm{h}}(t)=\begin{cases} cR_0^{-1}H_0^{-1}(2q_0-1)^{-1/2}R(t)\cos^{-1}[1-q_0^{-1}R_0^{-1}(2q_0-1)R(t)], & \\ & q_0>\tfrac{1}{2} \quad (k=1), \\ 2cH_0^{-1}[R(t)/R_0]^{3/2}, & q_0=\tfrac{1}{2} \quad (k=0), \\ cR_0^{-1}H_0^{-1}(1-2q_0)^{-1/2}R(t)\cosh^{-1}[1-q_0^{-1}R_0^{-1}(1-2q_0)R(t)], & \\ & q_0<\tfrac{1}{2} \quad (k=-1). \end{cases} \tag{4.75}$$

It can be shown that in the limit of small t, for the early epoch of the matter-dominated era, one gets the following expression for d_{h}:

$$d_{\mathrm{h}}(t)\to cH_0^{-1}(q_0/2)^{-1/2}(R/R_0)^{3/2}\approx ct/3. \tag{4.76}$$

Here $R(t)$ is much smaller than R_0. From (4.75) it is clear that for $q_0\leqslant\frac{1}{2}$, $R(t)$ increases without limit as t tends to infinity, so that $d_{\mathrm{h}}(t)$ increases faster than $R(t)$ and the particle horizon will eventually include all comoving particles, given sufficient time. For $q_0>\frac{1}{2}$, the universe is spatially finite, with a circumference given by

$$L(t)=2\pi R(t). \tag{4.77}$$

(See the discussion following (3.25).) At any time t we can see a fraction of this circumference given by (4.13) and (4.75) as follows:

$$d_{\mathrm{h}}(t)/L(t)=(2\pi)^{-1}\cos^{-1}[1-q_0^{-1}R_0^{-1}(2q_0-1)R(t)]. \tag{4.78}$$

Comoving particles within this fraction are visible. When $R(t)$ reaches its maximum value given by (4.21), this fraction will be $\frac{1}{2}$, and we shall see all the way to the 'antipodes'. This fraction remains less than unity until $R(t)$ reaches zero again, so we shall not be able to see all the way around the universe until that happens. If $q_0=1$ and $H_0^{-1}=13\times10^9$ years, the present circumference is 80×10^9 light years and the particle horizon is at 20×10^9 light years.

There may be some events in some cosmological models that we shall never see. It is clear from (4.72) that an event that occurs at time t_1 at the coordinate value r_1 will become visible at $r=0$ at a time t given by (4.72). If the t' integral diverges as t tends to infinity (or at a time that R reaches zero again), then it will be possible to receive signals from any event. However, if the t' integral converges for large t then we can receive signals from only those events for which

$$\int_0^{r_1}\mathrm{d}r/(1-kr^2)^{1/2}\leqslant c\int_{t_1}^{t_{\max}}\mathrm{d}t'/R(t') \tag{4.79}$$

where $t_{\max}$ is either infinity or the time of the next contraction: $R(t_{\max}) = 0$. This is referred to by Rindler as an *event horizon*. It is readily verified that for $q_0 < \frac{1}{2}$ or $q_0 = \frac{1}{2}$, the t' integral diverges as t tends to infinity so that there is no event horizon. For $q_0 > \frac{1}{2}$, $t_{\max} = 2T$, where T is given by (4.21). In this case an event horizon exists and the only events occurring at time t_1 that will be visible before R reaches zero again are those within a proper distance $d_E(t_1)$ given as follows:

$$\begin{aligned} d_E(t_1) &= cR(t_1) \int_{t_1}^{t_{\max}} \mathrm{d}t'/R(t') \\ &= cR_0^{-1}H_0^{-1}(2q_0 - 1)^{-1/2} R(t_1)\{2\pi - \cos^{-1}[1 - q_0^{-1}R_0^{-1}(2q_0 - 1)R(t_1)]\}. \end{aligned} \tag{4.80}$$

If $q_0 = 1$ and $H_0^{-1} = 13 \times 10^9$ years, then the only events occurring now that will ever become visible are those within a proper distance of 61×10^9 light years.

5

The Hubble constant and the deceleration parameter

5.1 Introduction

In the last two chapters we developed the mathematical framework, both kinematical and dynamical, to study various cosmological models that may represent, albeit as an idealization, the universe that we inhabit. In this chapter we discuss in some detail the observational aspects that must be considered to connect the models to reality. We will first give an account of earlier developments of this subject and mention more recent work towards the end of Section 5.4 and in the next chapter.

Two of the most important observational parameters in cosmology are the present values of Hubble's constant H_0 and the deceleration parameter q_0. Hubble's constant determines the present rate of expansion of the universe through the first term on the right hand side of (3.69). We write the following approximate form of this equation here again for convenience:

$$z = H_0 l/c + \tfrac{1}{2}(1 + q_0)H_0^2 l^2/c^2. \tag{5.1}$$

Thus in the limit of small distances the red-shift is given by H_0 times the distance divided by c. The deceleration parameter determines the rate at which the expansion is slowing down (or speeding up). As we see in (5.1), q_0 occurs in the second order term in a power series expansion in terms of l, the distance. Thus q_0 is determined by galaxies which are further than the ones from which H_0 is determined.

As we saw in the last chapter in the case of the Friedmann models, the parameters H_0 and q_0 determine these models completely. For example, if no pressure exists, the age of the universe t_0 is given in terms of H_0, q_0 by (4.18). We then get three possibilities. In the cases $k = 1, 0, -1$ we get $q_0 > \frac{1}{2}$, $q_0 = \frac{1}{2}$ and $q_0 < \frac{1}{2}$ respectively. In these cases the age of the universe is given respectively by (4.23), (4.25) and (4.31). Thus if we knew all three quantities

H_0, q_0, t_0 precisely, we could, in principle, decide which of the three models is correct, assuming, of course, zero pressure and other implicit assumptions that go into the definition of these models. We could then know all the large-scale physical properties of the universe.

Although, in principle, the determination of H_0 and q_0 is straightforward, in practice many difficulties arise and in this chapter we will consider some of these difficulties. This chapter is based mainly on the reviews by Sandage (1970, 1987), Gunn (1978), Longair (1978, 1983) and Bagla, Padmanabhan and Narlikar (1996).

5.2 Measurement of H_0

As mentioned earlier H_0 is measured from 'local' galaxies which are relatively nearby, whereas q_0 requires consideration of more distant galaxies. The first complication in measuring H_0 is that galaxies possess random motion of the order of 200 km s^{-1} which is caused by local gravitational perturbation, or 'lumpiness' of the galactic distribution, on a scale of about two million light years, which is the size of a small cluster of galaxies. For a large cluster which has rotational velocity about some centre, this random motion can be much higher. One can take account of this random motion, but for this one has to take a large sample. Secondly, there may be local anisotropy, but on quite a large scale, which may distort the velocity field in some directions for red-shifts which imply velocities smaller than about 4000 km s^{-1}. This anisotropy may arise partly due to an abnormal concentration of groups of galaxies such as the Virgo cluster on a scale of about 30 million light years.

Another complication in the measurement of H_0 is the rotational motion of the Sun about the galactic centre of the Milky Way, which amounts to approximately 300 km s^{-1} in the direction of Cygnus. This velocity is an appreciable fraction of the recessional velocities of nearby galaxies in the direction of Cygnus, so this effect appears as an added anisotropy in the observed velocity field. To map this velocity field precisely, accurately subtracting any spurious velocities, requires data from the Southern Hemisphere, which have only recently been forthcoming.

Thus an accurate measurement of H_0 requires precise distance determinations of nearby objects. Distance calibration is a stepwise process in which errors proliferate at each step. First one measures the apparent brightness of well-known objects in nearby galaxies, which can be resolved optically. If the absolute brightnesses of these objects are known from another source, the distance can be determined by the inverse-square law

of the falling of the intensity. Because the absolute luminosities can be related to the periods of Cepheid variable stars, these stars are excellent indicators of distance.

The term Cepheid variables derives from a particular member of this class known as Delta Cephei. In the early part of this century H. S. Leavitt and H. Shapley found a relationship between the observed period of variation of the Cepheids and their intrinsic brightness. In 1923 Hubble was for the first time able to resolve the nearby galaxy Andromeda into separate stars, and locate Cepheid variables in it. Using the Leavitt–Shapley relation he concluded that the Andromeda nebula was at a distance of 900 000 light years, which was clearly outside our galaxy, since it was more than ten times further than the most distant object known in our galaxy. Later, however, Baade (1952) and others showed that there are, in fact, two types of Cepheid variables, and that those that Leavitt and Shapley observed and those that Hubble observed belong to the two different types, so that Hubble used the wrong period–luminosity relation. The distance to the Andromeda nebula turns out to be over two million light years. (See Figs 5.1 and 5.2 for further information about Cepheid variables.)

The distance range over which H_0 can be determined is not very large. It starts from about 10^7 light years, which is far enough so that recessional velocities begin to dominate the random velocities, and ends at about 6×10^7 light years, which is the upper limit for the distance indicators to be resolved by powerful optical telescopes. There are various possible distance indicators in this range, such as red and blue supergiants, the angular size of HII regions, normal novae and possibly supernovae. The nearer of these are first calibrated with Cepheid variables and then used in turn as more distant indicators.

Because of Hubble's error alluded to above, the value of H_0 for more than a decade following 1936 was taken to be about 165 km s^{-1} per million light years or about 538 km s^{-1} Mpc^{-1}. In the simplest cosmological models this meant an age of the universe of only 1.8×10^9 years. Even in the 1930s this was known to violate the age of the Earth as known from geological studies, such as the age of the Earth's crustal rocks and the lower limit of 7×10^9 years for the age of the Earth's radioactive elements.

There was a controversy in the 1930s and 1940s as to whether the value of H_0 was wrong, or the Friedmann models. Lemaître and Eddington, for example, devised models with a 'cosmological constant' (about which we will learn more in the next chapter) to fit the high value of H_0. The controversy was finally settled in the 1950s following the work of Baade cited earlier, which started a detailed recalibration of the period–luminosity

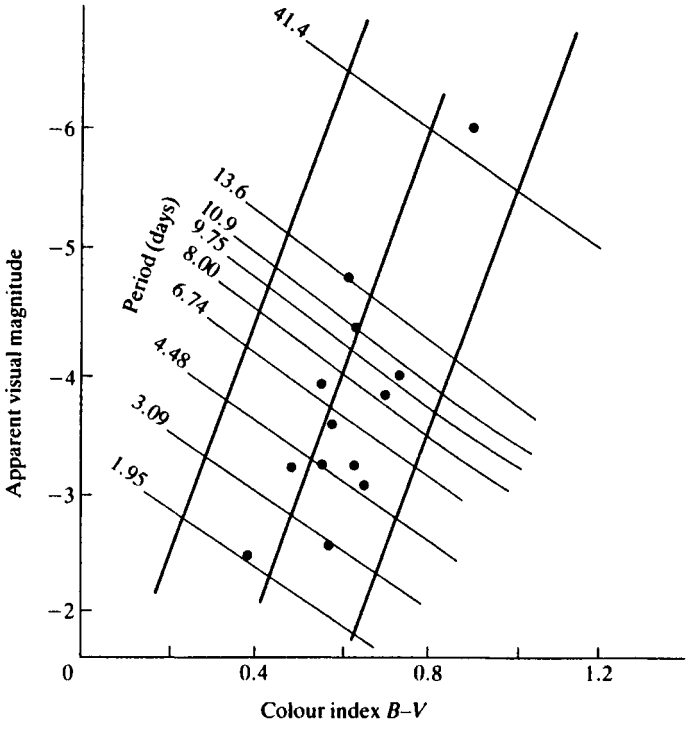


Fig. 5.1. In this diagram the visual luminosity M_V of a star is plotted against its colour $B–V$ which is a measure of its temperature. Here $B–V = 2.5 \log(l_V/l_B) - \text{constant}$, where l_V, l_B are the luminosities integrated over the visual and blue ranges of the spectrum, respectively. Pulsating stars lie in the region between the two outer lines with positive slope. The stars pulsate with periods that increase with increasing luminosity. Cepheids of the same period can differ in absolute luminosity by one magnitude, the bluer Cepheids being brighter.

relation of Cepheid variables, to which contributions were made by Kraft (1961), Sandage and Tammann (1968, 1969) and others. High-precision photometric methods developed in the 1950s by Eggen, Johnson and others also contributed to this progress. These improved calibrations, and the precise distance determinations in the crucial range for H_0 mentioned earlier, such as some highly resolved systems centred on the giant spiral M81, have considerably improved the measurement of H_0. There is, however, still an uncertainty, the present range of values being $15 \leq H_0 \leq 30$ km s^{-1} 10^{-6} (light year)$^{-1}$ or about $50 \leq H_0 \leq 100$ km s^{-1} Mpc^{-1}. This makes the age of the universe from 13 to 20 billion years approximately. Among those who have contributed to this new determination of H_0 are

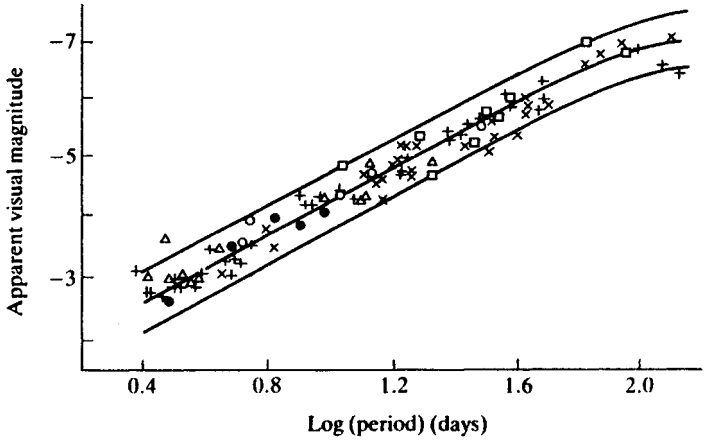


Fig. 5.2. In this diagram the apparent visual magnitude is plotted against the logarithm of the period (in days). The scatter in this diagram is caused by variation of the colour indicated in Fig. 5.1. An accurate calibration can be made if this *P–L* (period–luminosity) diagram is considered in conjunction with the colour of the Cepheids. Calibrating Cepheids are the galactic cluster Cepheids (solid circles) and the *h* and Perseus Association (open circles). Other symbols represent Cepheids belonging to the Local Group of galaxies.

Sandage and Tammann (1975), de Vaucouleurs (1977), Tully and Fisher (1977) and Van den Bergh (1975).

5.3 Measurement of q_0

If one had knowledge of 'standard candles', that is, objects of fixed, known absolute luminosities, then the apparent luminosities of these objects would be a measure of their distance, and by determining their red-shifts one could plot a graph of red-shift versus apparent luminosity, from which, in principle, one could read off the values of H_0 and q_0. The apparent luminosity of a source is usually described by its so-called bolometric magnitude, denoted by m_{bol}. For small red-shifts z the following relation obtains between m_{bol} and z (see the Appendix to this chapter for a definition of m_{bol} and a derivation of this relation, (5.17) below):

$$m_{\text{bol}} = 5 \log_{10}(cz) + 1.086(1 - q_0)z + \text{constant}. \tag{5.2}$$

The constant contains H_0 (see the Appendix, p. 90). This relation is true for all Friedmann models, but for small z.

Equation (5.2) is useful because it relates quantities m_{bol} and z which are

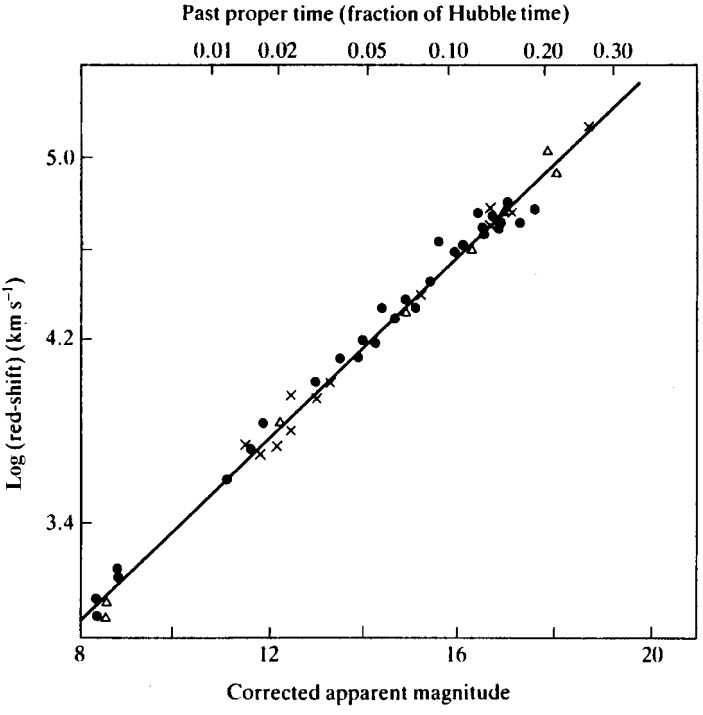


Fig. 5.3. This is a Hubble diagram for 42 galaxies in clusters (see paragraph following Equation (5.2)). Triangles represent non-radio sources measured by W. A. Baum. Crosses represent radio sources and closed circles represent other non-radio sources. These were measured by the 200 in telescope at Mount Palomar (Sandage, 1970).

directly measurable. All observations confirm the leading term in (5.2). Figure 5.3 gives one of these observational plots. The figure has data for 42 clusters of galaxies, each of which has a good distance indicator, which is the brightest in the cluster. The small horizontal dispersion about the line, with the theoretical slope of 5, shows the near constancy of absolute luminosity for galaxies chosen this way. It is clear, however, from (5.2) that the value of q_0 cannot be determined from the data in Fig. 5.3, and that one must resort to much higher red-shifts. The data of Fig. 5.3 extend only till $z = 0.46$, and for this kind of red-shift any significant variation in q_0 which could decide between different models gives a variation in m_{bol} which is equivalent to the scatter of galaxies about the mean line, and for this reason not very useful. Further, there are uncertainties in the various corrections to observed magnitudes which are themselves comparable to the

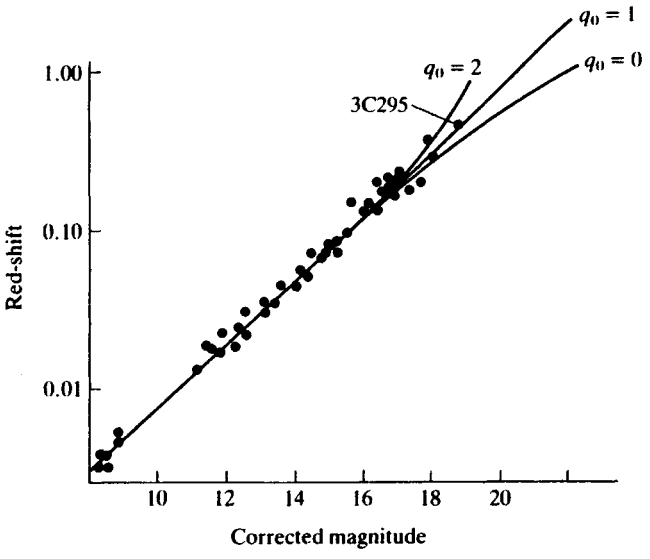


Fig. 5.4. This diagram is an idealized version of Fig. 5.3, showing the extrapolation to regions which would determine the value of q_0 (Weinberg, 1972).

variation. One also requires knowledge of the way absolute luminosities evolve during the period which has elapsed since light left those distant galaxies, due to evolution of their stellar content.

Figure 5.4 displays an idealized version of Fig. 5.3, and shows clearly that one needs galaxies with higher red-shifts to distinguish between different values of q_0. The 'test objects' in Fig. 5.3 are giant elliptical galaxies, which tend to be the brightest galaxies in any cluster, and have similar light distribution curves, that is, curves which give a plot of intensity versus wavelength or frequency.

As mentioned earlier, several difficulties arise when one attempts to measure q_0 accurately. One of these is that galaxies do not have well-defined boundaries, so the intensity depends to some extent on the aperture with which one measures it. For this reason all measurements have to be corrected to some adopted 'standard galaxy diameter'.

Every galaxy has an intrinsic frequency distribution of light, that is, an intensity–frequency plot. For distant galaxies this frequency distribution is distorted, because their visual or blue magnitudes reflect their absolute luminosities at higher frequencies than for near galaxies. Thus the left hand side of (5.17) below is replaced by $m - M - k(z)$, where $k(z)$ is an

explicitly known function of z, calculated by Oke and Sandage (1968), known as the k term. In the earlier alternative procedure due to Baum (1957), the luminosity distribution is measured directly for each galaxy and no k term is needed.

Our galaxy absorbs a certain amount of radiation coming from objects outside the galaxy. Considering the galaxy to be a flat slab, the distance through which light must travel in the galaxy on its way to the observer is proportional to cosec b, where b is the angle between the line of sight and the plane of the galaxy. Due to absorption in the galaxy the light will thus be decreased in intensity by a factor $\exp(-\lambda \operatorname{cosec} b)$, where λ is a constant which can be determined from some known extragalactic objects. The distance modulus in (5.17) will then be corrected as follows:

$$(m-M)_{\text{corr}} = m - M - k(k) - A(b), \tag{5.3}$$

where we have approximately, $A(b) = 0.25 \operatorname{cosec} b$. This is a somewhat simplified description of the correction due to absorption by the galaxy.

There are still uncertainties in the precise determination of the absolute luminosities of the brightest E galaxies (giant ellipticals). Any change in the estimated distance to nearby objects such as the Hyades or the Virgo cluster would require a corresponding change in these absolute luminosities.

If there is no definite upper limit to the absolute luminosity of a cluster of galaxies, then there would be a tendency to select richer clusters at greater distances resulting in a slight increase of the absolute luminosities of the brightest galaxies with increasing z. This is known as the 'Scott Effect' and may result in a slight overestimation of q_0 but would not have a significant effect on H_0.

The rotation of the Sun about the galactic centre and the existence of a local anisotropy in the galactic velocity field have already been mentioned. The evolutionary effects which were mentioned briefly will be considered in more detail in the Appendix.

The observation and analysis of radio sources have played a significant part in cosmology. For reviews of this topic we refer to Longair (1978, 1983, 1998). One of the most important applications of radio astronomy has been the detection and identification of quasars, which are powerful emitters in the radio band. The quasars 3C48 and 3C273, for example, were also identified through optical telescopes and they appeared to be stars, but with peculiar emission lines. These seemed peculiar because the objects were thought to be stars within our galaxy. It was realized later that the emission lines were familiar ones which had been red-shifted by the

equivalent of $z = 0.367$ and $z = 0.158$ respectively, so that these were at distances of 5 and 3 billion light years respectively. Many other quasars have since been discovered; the quasar 3C9, for example, has a red-shift of $z = 2.012$. Since the quasars are so bright at such distances, their energy output must be enormous, especially because this energy comes from regions which are only a few light days or weeks across. This follows from the fact that the brightness of the quasars varies substantially over periods of days or weeks. How this enormous energy output is possible from such a small region has been a puzzle for a long time. One of the reasonably successful models is that of a large black hole at the centre of the galaxy which swallows stars, which in the process get disrupted by tidal gravitational forces and give off large amounts of radiation. Such a process can account for the energy output of quasars provided this enormous output does not last for more than a few tens of millions of years at most. There is evidence that quasars do, in fact, only last for a few tens of millions of years.

An intriguing aspect of the quasar problem, which seems worth pursuing carefully, is the fact that there seems to be a cut-off in quasar red-shifts at about $z = 4$. For about ten years the highest quasar red-shift known was $z = 3.53$, although techniques had improved so that higher red-shifts could have been observed. According to M. Smith (see Longair (1983)), there are seven quasars in the Hoag and Smith survey in the red-shift range $2.5 < z < 3.5$, and so eight or nine of them should have been detectable in the Osmer deep survey carried out later, with $3.5 < z < 4.7$, provided the comoving spatial density of quasars remains constant. In fact none was found, although one larger red-shift quasar is now known. However, the question is a statistical one and a great deal more work has to be done before any definite conclusion can be drawn. If indeed there is a cut-off in quasar red-shifts around $z = 4$, the following reasons might be adduced for this phenomenon:

(*a*) There might be intervening dust in the discs of galaxies so that by the time one gets to distances corresponding to a red-shift of about 3.5, a substantial portion of the celestial sphere might be covered by these discs.

(*b*) The most prominent emission line through which quasar red-shifts are observed is the Lyman-α line. It is possible that there may be a lack of continuum photons or gas around large red-shift quasars which inhibits the Lyman-α lines.

(*c*) It is possible that it takes a long time for the black holes, which are at the centre of the largest quasars, to grow. There may be quasars

with z much larger than 4, but they may not have grown to the hyperluminous stage.

(*d*) The dust and gas in the intervening young galaxies may absorb a significant part of the emission from quasars and reradiate in the infrared band.

(*e*) There is the intriguing possibility that there are no galaxies beyond about $z = 4$, because galaxies may condense out of the intergalactic gas until about $z = 4$.

From the above considerations it would appear that there used to be much more violent activity in the universe at red-shifts of about 2–4 than there is now. This does indicate evolution of the universe and is consistent with the existence of the cosmic background radiation.

Radio astronomy has provided a valuable additional approach to observational cosmology. One of the reasons for its importance is that numerous faint radio sources have been detected, many of which lie presumably at great distances, which have not been optically identified and probably cannot be so identified, at least in the foreseeable future. However, the red-shifts of these sources are for this reason not known, so that one has to follow a programme other than the Hubble programme (outlined above) to elicit information from these faint sources that may be of cosmological interest. Such a programme is that of number counts, in which one determines the number of sources as a function of flux density. It can be shown that in a uniform Euclidean world model, the number of sources N whose flux density is greater than S is proportional to $S^{-3/2}$. By observing and plotting the departure of the actual distribution from this law one can get information about the correct model of the universe. Although there are many uncertainties, some interesting points have emerged. For example, there is evidence that there have been significant variations in the population of radio sources with cosmic epoch. We refer to Weinberg (1972) and Longair (1983, 1998) for more details.

5.4 Further remarks about observational cosmology

The distant galaxies that are used for the measurement of q_0 are all in clusters. In this case a substantial proportion of the mass is not in the galaxies, but is distributed smoothly between the galaxies. A galaxy moving through this stuff – whatever form it has – experiences so-called *dynamical friction* (Chandrasekhar, 1960), which is a kind of frictional drag which the moving galaxy experiences by virtue of the high-density gravitational

wake behind it. This effect on clusters of galaxies was first studied by Ostriker and Tremaine (1975). The net effect of this is that galaxies which are near the main one are swallowed up by it and its luminosity is thereby increased. The final effect of this on the value of q_0 is somewhat uncertain. Although the brightness of the cannibal galaxy increases, it becomes extended and of low density. Since the luminosity of galaxies is measured in a fixed aperture, it is not clear if the luminosity increases or decreases. The situation is rather complex and a great deal of theoretical and observational work has to be done before this process is fully understood. We refer the reader to Gunn (1978) for further material on this. As is clear from Gunn's article, one of the most important problems is to determine precisely the evolutionary effects on galaxies, clusters of galaxies and quasars.

From (4.14) we recall that if the pressure is negligible, the ratio of the present density to the critical density is twice the deceleration parameter. This ratio is usually denoted by Ω_0 and referred to as the density parameter. The galactic mass density of the universe, that is, the mass of visible matter, is of the order of about a tenth or less of the critical density, so that Ω_0 is around 0.1 or less. However, although there is much uncertainty in the observed value of q_0, indications are that it is about unity or a bit less. There is thus a disagreement between observations and (4.14). This discrepancy has been a long-standing problem in cosmology, and various explanations have been put forward for it. One possibility is that the value of q_0 obtained so far is higher than it should be, because of evolution or selection effects. In fact, it is known that if one determines q_0 solely on the data from quasars, one gets a value somewhat higher than unity. However, if one assumes for the present that q_0 is indeed of order unity, then (4.14) implies that the density of the universe is about 2×10^{-29} g cm^{-3}. This is an order of magnitude or more higher than that observed. Thus there may be some 'missing mass' which is not directly observable. One possibility is that the missing mass resides in the intergalactic space in clusters of galaxies. If a cluster is gravitationally bound, then by the use of the virial theorem one can estimate its mass, which turns out to be several times higher than would be obtained by adding the masses of individual galaxies (see, for example, Karachentsev (1996)). If this is the case for all or most clusters, the density of the universe would be raised considerably. However, although, for example, the Coma cluster appears to be bound (there is no certainty of this), others like those in Virgo or Hercules are highly irregular and may not be bound.

The missing mass may reside in the space between clusters of galaxies.

The total volume outside clusters is approximately 500 times the volume within clusters, so that even a density between clusters which is one-tenth of that within clusters would add significantly to the average density. If indeed there is mass between clusters, presumably it is in a form which does not radiate significantly in the visible spectrum such as atomic or ionized hydrogen, dwarf galaxies which are very faint, black holes, etc. It is uncertain if these or other forms of matter exist in the intergalactic space.

The missing mass may also reside in highly relativistic particles such as cosmic rays, photons, neutrinos or gravitons. These may either be relics from the early universe (see Chapter 8), or may be created in various processes in more recent times. As regards the cosmic background radiation, from the fact that its temperature is 2.7 K and the Stefan–Boltzmann law one can deduce that the associated energy density is about 4.4×10^{-34} g cm^{-3}. The density of cosmic rays and other known forms of radiation is much less than this. As regards 'cosmic background neutrinos', the temperature of these would be $(\frac{4}{11})^{1/3}$ times the temperature of the cosmic background radiation (see Chapter 8 and Weinberg (1977)), or about 2 K. Assuming that the number density of these neutrinos is the same as that of photons, that is, approximately 10^9 for each baryon, this would not make a significant contribution to the overall density if the neutrinos are massless. However, in recent years there have been indications that neutrinos may have a non-zero but small mass, of the order of a few electron volts (recall that the electron has a mass of about half a million electron volts). Thus if neutrinos had a mass of 10 eV, the contribution of neutrinos to the density would be about ten times that due to the visible matter in the universe. However, there are, as usual, many uncertainties in this analysis, and one must wait for more accurate data and theories (see, for example, Tayler (1983)). One point of some interest is that if the density is dominated by massless particles the equation of state becomes that of pure radiation (see Section 4.3) and instead of (4.14) we get $\varepsilon_0/\varepsilon_c = q_0$, and the density required for a given q_0 and H_0 is half of that needed for a zero pressure model.

In the rest of this section we remark on more recent work, following the important review by Sandage (1987). As is clear from the foregoing discussion, one of the most important problems in observational cosmology is the determination of q_0 by comparing (5.2) with observations. Some of the difficulties have already been mentioned in the discussion of Fig. 5.3. Sandage refers to this problem as the '$m(z)$ test'. Following many years of work by various people (Sandage, 1968, 1972a,b, 1975a,b; Sandage and Hardy, 1973; Gunn and Oke, 1975; Kristian, Sandage and Westphal, 1978; Sandage and Tammann, 1983, 1986), Sandage feels that the $m(z)$ test is

inconclusive mainly because of uncertainty in the evolution of standard candles. It is clear from (5.2) (see Equation (15) of Sandage (1987)) that for small z, $\log z = 0.2\, m + \text{constant}$. This is indeed indicated by observations until about $z = 0.5$. There are departures from this linear relation between $\log z$ and m for $z > 0.5$. According to Sandage, there may be three different reasons for this departure, as follows: (*a*) a value of q_0 in the range $0.5 < q_0 < 2$, (*b*) genuine small departures from linearity for small z, and (*c*) the combined effects of $q_0 \neq 1$ and evolution of luminosity (see (5.20) below). The reader is referred to studies of the $m(z)$ relation by Lilley and Longair (1984), Lilley, Longair and Allington-Smith (1985) and Spinrad (1986) for more material on this question.

Gross deviations from the $m(z)$ relation – the latter referred to sometimes as the 'Hubble flow' – although sometimes claimed (Arp, 1967, 1980; Burbidge, 1981), have not been substantiated. Large perturbations to the Hubble flow connected with the Local Group of galaxies may, however, exist (see, for example, Davies *et al.*, 1987).

The angular diameter θ of some standard objects also has a dependence on z and q_0 and can be used as a test, as was first suggested by Hoyle (1959). Strictly speaking, the $m(z)$ relation and $\theta(z)$ relation should be derivable from each other, but as θ is directly measurable, this can provide a useful additional check. There are uncertainties in the $\theta(z)$ programme; one has firstly, to give a precise definition of angular diameter (for example, angular size to a given isophote, that is, a contour of equal apparent brightness) that will be valid for sources of all magnitudes and, secondly, there are the difficulties of evolutionary and selection effects similar to those for the $m(z)$ test (see, for example, Sandage (1972a), Djorgovski and Spinrad (1981)). When discussing the $\theta(z)$ programme Sandage makes an important point about observational cosmology. There are essentially three tests, namely the $m(z)$, $\theta(z)$ and $N(m)$ tests, where $N(m)$ is the number of galaxies brighter than the apparent magnitude m. Sandage thinks that the predictions of the Friedmann models are not confirmed in detail by any of these tests 'using the data as they are directly measured'. To get agreement one usually invokes evolutionary effects with time. This would be justified only if one had independent evidence that the standard model is correct, which is, in fact, the object of the exercise.

An important difficulty is that of selection effects, which, roughly speaking, means that in a sample of sources of limited flux (apparent magnitude), the average absolute luminosity of the nearby members is, in general, less than that of more distant members. Selection effects can cause serious uncertainties, such as in the determination of the value of

H_0. The reader is referred to Sandage (1972c), Sandage, Tammann and Yahil (1979), Spaenhauer (1978), Tammann *et al.* (1979) and Kraan-Korteweg, Sandage and Tammann (1984) for more material on selection effects, particularly in the form known as the Malmquist bias.

An important observational problem is large scale clustering of galaxies ('superclusters'), first suggested by Hubble (1934). Hubble concluded that the universe was homogeneous on the largest scale that he could measure (to a depth of $m \sim 22$), but that it was clumped or clustered on an intermediate scale. A study by Crane and Saslaw (1986) obtained similar results and drew the same conclusions as Hubble. As regards intermediate structure, an important discovery of the 1980s has been that of 'filaments' along which galaxies tend to concentrate, initially noticed by Peebles and his collaborators (see, for example, Seldner, Siebers, Groth and Peebles (1977)). Unusually large empty regions ('voids') have also been detected. Much work has been done on this matter and is continuing; see for example, studies by Tarenghi *et al.* (1979), Gregory, Thompson and Tifft (1981), Kirshner, Oemler, Schechter and Schectman (1981), Gregory and Thompson (1982), Chincarini, Giovanelli and Haynes (1983), Huchra, Davis, Latham and Tonry (1983), and the review by Oort (1983). (See especially the recent book by Saslaw, 2000.)

The time scale test, one in which one compares the age of the universe from observations and models, has been mentioned earlier. The main uncertainty here is the value of H_0, which varies by a factor 2. For $H_0 = 50$ km s^{-1} Mpc^{-1}, the age is about 19.5×10^9 years, whereas for $H_0 = 100$ km s^{-1} Mpc^{-1} one gets approximately 9.8×10^9 years. The comparison with observation is somewhat inconclusive (Sandage, Katem and Sandage, 1981; Sandage, 1982).

Sandage suggests the following programme for observation cosmology for the next two decades (writing in 1987). This is a succinct version of the description of the programme given by Sandage (1987).

(*a*) Proof or otherwise that the red-shift represents a true expansion of the universe.
(*b*) Proof or otherwise of evolution of galaxies in the look-back time.
(*c*) Comparison of the value of H_0 with that obtained from the globular cluster time scale. (The globular clusters are among the oldest objects in the Galaxy.)
(*d*) The compatibility of clustering properties of galaxies with possible variations of the Hubble flow.
(*e*) Studies of the galaxy luminosity functions for different types of galaxies (see (5.18), (5.19) below).

(f) The detection of $\delta T/T$ fluctuations in the temperature T of the cosmic background radiation at a level of one part in $\sim 10^{5.5}$ on small angular scales. This would have an important bearing on galaxy formation.

Appendix to Chapter 5

In this Appendix we derive the formula (5.2), and give some relevant definitions. For more details we refer to Weinberg (1972). The absolute luminosity L of a source is the amount of radiation emitted by the source per unit time. The apparent luminosity l' is the amount of radiation received by the observer per unit time per unit area of the telescopic mirror or plate. In Euclidean space, the apparent luminosity of a source at rest at a distance d would be $L/(4\pi d^2)$, by the usual inverse square law of the decrease of radiation. By analogy with this one defines a *luminosity distance* d_L in the more general case as follows:

$$d_L = (L/4\pi l')^{1/2}. \tag{5.4}$$

By taking into account the red-shift of the moving source one can show that in the general case the apparent luminosity is related to the absolute luminosity as follows:

$$l' = LR^2(t_1)/4\pi R^4(t_0) r_1^2. \tag{5.5}$$

(See Equation (14.4.12) of Weinberg (1972).) Here the source is at the coordinate radius r_1, the times t_0 and t_1 being those of the reception and emission of the radiation. Equation (5.5) is valid for all three values of k. From (5.4) and (5.5) we get

$$d_L = R^2(t_0) r_1 / R(t_1). \tag{5.6}$$

By generalizing (3.53) to the two other values of k we get

$$c\int_{t_1}^{t_0} dt/R(t) = \int_0^{r_1} dr/(1-kr^2)^{1/2} = f(r_1), \tag{5.7}$$

where $f(r_1) = \sin^{-1} r_1$, r_1, $\sinh^{-1} r_1$ according to whether $k = 1, 0, -1$.

Next we write (3.68), with $t = t_1$:

$$z = (t_0 - t_1)H_0 + (t_0 - t_1)^2(\tfrac{1}{2}q_0 + 1)H_0^2 + \cdots. \tag{5.8}$$

Inverting this power series, we get (see Fig. 5.5)

$$t_0 - t_1 = H_0^{-1} z - H_0^{-1}(1 + \tfrac{1}{2}q_0)z^2 + \cdots. \tag{5.9}$$

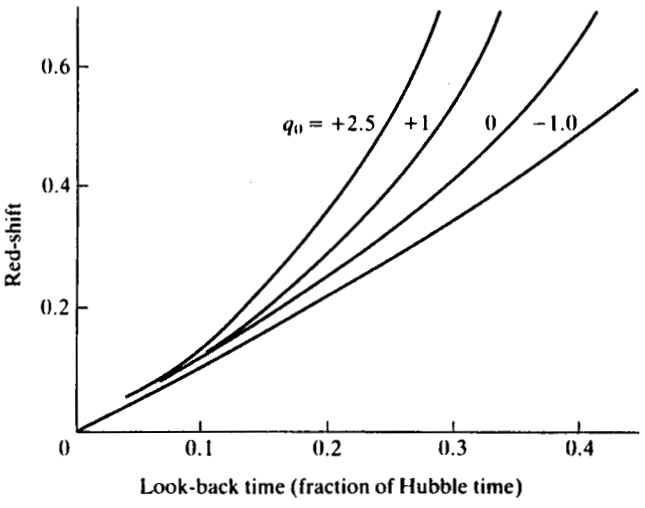


Fig. 5.5. In this diagram Equation (5.9) is illustrated. The 'look-back time' is $t_0 - t_1$ of (5.9) (Sandage, 1970).

With the use of (3.64) or the right hand side of (3.66) in (5.7) we get

$$r_1 + O(r_1^3) = cR_0^{-1}[t_0 - t_1 + \tfrac{1}{2}H_0(t_0 - t_1)^2 + \cdots]. \tag{5.10}$$

With the use of (5.9) and (5.10) we get r_1 in terms of the red-shift:

$$r_1 = c(R_0H_0)^{-1}[z - \tfrac{1}{2}(1 + q_0)z^2 + \cdots]. \tag{5.11}$$

Equation (5.6) then gives d_L as a power series in z:

$$d_L = cH_0^{-1}[z + \tfrac{1}{2}(1 - q_0)z^2 + \cdots]. \tag{5.12}$$

This can be transformed to a formula for the apparent luminosity l':

$$l' = L/(4\pi d_L^2) = c^{-2}(LH_0^2/4\pi z^2)[1 + (q_0 - 1)z + \cdots]. \tag{5.13}$$

The apparent luminosity l' is usually expressed in terms of an *apparent bolometric magnitude* m_{bol}, or simply m, which is defined as follows:

$$l' = 10^{-2m/5} \times 2.52 \times 10^{-5} \text{ erg cm}^{-2} \text{ s}^{-1}. \tag{5.14}$$

The *absolute bolometric magnitude* M is defined as the apparent bolometric magnitude the source would have at a distance of 10 pc:

$$L = 10^{-2M/5} \times 3.02 \times 10^{35} \text{ erg s}^{-1}. \tag{5.15}$$

Thus the *distance modulus* $m - M$ can be defined as follows:

$$d_L = 10^{1+(m-M)/5}\text{pc}. \tag{5.16}$$

Equations (5.12) and (5.16) can now be combined to give the desired relation between the distance modulus (or the bolometric magnitude) and the red-shift:

$$\begin{aligned} m - M = {} & 25 - 5\log_{10}H_0(\text{km s}^{-1}\ \text{Mpc}^{-1}) \\ & + 5\log_{10}(cz)\ (\text{km s}^{-1}) + 1.086(1-q_0)z + \cdots . \end{aligned} \tag{5.17}$$

The apparent magnitudes m_{U}, m_{B}, etc., in the ultraviolet, blue, photographic, visual (see Fig. 5.1), and infrared wavelength bands are defined similarly to (5.15) and (5.16) but with different constants chosen so that all apparent magnitudes will be the same for stars of a certain spectral type and magnitude. The *colour index* is the quantity $m_{\text{B}} - m_{\text{v}} = M_{\text{B}} - M_{\text{v}}$.

We will now give a brief description of the correction to the deceleration parameter due to possible variation of the luminosity L with evolution of galaxies. As we observe distant galaxies, we are looking at earlier times when these galaxies were younger. It is possible that the luminosity of the brightest E galaxies is a function of the time t_1 at which the light was emitted: $L(t_1)$. We see from (5.9) that in this case the L in (5.13) should be replaced by the following expression:

$$\begin{aligned} L(t_1) &= L(t_0)[1 - E_0(t_0 - t_1) + \cdots] \\ &= L(t_0)[1 - E_0 z/H_0 + \cdots] \end{aligned} \tag{5.18}$$

where

$$E_0 = \dot{L}(t_0)/L(t_0). \tag{5.19}$$

Substituting this into (5.13) we readily see that the overall effect is to replace q_0 with q_0^{eff}, where

$$q_0^{\text{eff}} = q_0 - E_0 H_0. \tag{5.20}$$

There are many uncertainties in the value of E_0. Any value of E_0 of the order of $0.04/10^9$ years or above would have a significant effect on the value of q_0^{eff}. It is possible that E_0 is negligible.

We end this Appendix with some remarks about dimensions. It is straightforward to check the dimensions of any of the equations in this book, but the following discussion may help the novice. As usual we denote by L, M, T the dimensions of length, mass and time respectively (the T here is not to be confused with the temperature, which is denoted by

T elsewhere in the book). We write $[X]$ for the dimension of the quantity X, and denote by unity the dimension of a dimensionless quantity. The following relations are easy to verify:

$$[G] = M^{-1}L^3T^{-2},\ [c] = LT^{-1}, \tag{5.21a}$$

$$[R] = L,\ [\dot{R}] = LT^{-1},\ [\ddot{R}] = LT^{-2}, \tag{5.21b}$$

$$[z] = [r] = [q_0] = 1,\ [H_0] = T^{-1}. \tag{5.21c}$$

In (5.21b) and (5.21c) R, r are respectively the scale factor in the Robertson–Walker metric and the coordinate radius. Other similar relations can be derived readily. We will now choose a few equations at random and verify the dimensions of each side of the equations. Consider (4.1), of which the left hand side has dimension LT^{-2} (see (5.21b)). The quantity ε on the right hand side is energy density, that is, energy divided by volume. Since energy has dimension ML^2T^{-2}, we get

$$[\varepsilon] = ML^2T^{-2}/L^3 = ML^{-1}T^{-2}. \tag{5.22}$$

This is the same as the dimension of p, the pressure, which is force per unit area. Force has the dimension MLT^{-2}; dividing this by L^2, the area, yields $ML^{-1}T^{-2}$ as in (5.22). The right hand side of (4.1) thus has dimension

$$[G][\varepsilon][R/c^2] = (M^{-1}L^3T^{-2})(ML^{-1}T^{-2})(L/L^2T^{-2}) = LT^{-2}, \tag{5.23}$$

as required. Consider (4.24), of which the left hand side is clearly dimensionless. Since H_0 has the dimension T^{-1}, H_0t is also dimensionless. In (4.41) both sides have the dimension T^{-2}. It might be instructive for the reader without much experience of this matter to check in detail the dimension of each equation.

We will consider more recent developments in observational cosmology at the end of the next chapter after a discussion of the cosmological constant.

6
Models with a cosmological constant

6.1 Introduction

From (4.1) we see that if we want a *static* solution of Einstein's equations, that is, one in which $R=0$, we must have $\varepsilon+3p=0$, which is a somewhat unphysical solution, because, assuming the energy density to be positive, the pressure must be negative. If we demand that the pressure be zero, then the energy density turns out also to be zero.

When Einstein formulated the equations of general relativity in 1915 the expansion of the universe had not been discovered, so that the possibility that the universe may be in a dynamic state did not occur to people. It was natural for Einstein to look for a *static* solution to his cosmological equations. But for the reasons mentioned above such a solution did not appear to exist. Einstein therefore modified his equations by adding the so-called 'cosmological term' to his equation (2.22), as follows:

$$R_{\mu\nu}-\tfrac{1}{2}g_{\mu\nu}R-\Lambda g_{\mu\nu}=\frac{8\pi GT_{\mu\nu}}{c^4}, \tag{6.1}$$

where Λ is the *cosmological constant*. Equations (3.76a) and (3.76b) are then modified as follows (note that $[\Lambda]=L^{-2}$):

$$3(\dot{R}^2+c^2k)=8\pi G\varepsilon R^2/c^2+c^2\Lambda R^2, \tag{6.2a}$$

$$2R\ddot{R}+\dot{R}^2+kc^2=-8\pi GpR^2/c^2+c^2\Lambda R^2. \tag{6.2b}$$

If we now demand a static solution with $R(t)=R_0$, a constant, and, say, zero pressure, we get the following values:

$$\varepsilon=(c^4\Lambda/4\pi G),\ k=\Lambda R_0^2. \tag{6.3}$$

Thus Λ must be positive and, correspondingly, we must choose $k=1$, so that the universe has positive spatial curvature. This is Einstein's static

universe. In later years Einstein regretted adding the cosmological term, because if he had been sure that the universe conformed to his original equations, the fact that no reasonable solutions exist representing a static universe would have led him to infer that the universe is in a dynamic state. He would still not have known if the universe is expanding or contracting, but the discovery of a dynamic state would have been an important one.

Apart from the static solution mentioned above, there are, of course, many dynamic solutions with the cosmological constant. These models were first studied by Lemaître so they are known as Lemaître models. In recent years other motivations have been found for introducing a cosmological term and such a term arises in many different contexts. We shall consider some of these later in this chapter and in other chapters. Introducing the cosmological term is like introducing a fictitious 'fluid' with energy–momentum tensor $T'_{\mu\nu}$ given by

$$T'_{\mu\nu} = (\varepsilon' + p')u_\mu u_\nu - p' g_{\mu\nu} = (8\pi G/c^4)^{-1} \Lambda g_{\mu\nu}, \tag{6.4}$$

so that the energy density and pressure of this fluid are given by $\varepsilon' = (c^4\Lambda/8\pi G)$, $p' = -(c^4\Lambda/8\pi G)$. For then (6.1) can be written as follows:

$$R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R = (8\pi G/c^4)(T_{\mu\nu} + T'_{\mu\nu}). \tag{6.5}$$

One can follow steps similar to those in Chapter 4 to derive the Lemaître models. Thus instead of (4.8) we get the following equation:

$$\varepsilon_c = 3c^2 H_0^2/8\pi G = -3kc^4/8\pi G R_0^2 + \varepsilon_0 + c^4\Lambda/8\pi G. \tag{6.6}$$

Recalling the density parameter Ω_0 introduced in Chapter 4 (see discussion following (4.9)), (6.6) can be written as follows:

$$c^2 k/R_0^2 H_0^2 = \Omega_0 - 1 + c^2\Lambda/3H_0^2. \tag{6.7}$$

Equation (4.10) is modified as:

$$\varepsilon_0 + 3p_0 = (3/4\pi G) q_0 H_0^2 c^2 + c^4\Lambda/4\pi G, \tag{6.8}$$

while (4.13) becomes

$$H_0^2(2q_0 - 1) = c^2 k/R_0^2 - \Lambda c^2. \tag{6.9}$$

Consider now the solutions that would obtain if we had zero pressure but non-zero Λ. It can be shown after some reduction, in which use is made of (6.7)–(6.9), and the fact that (3.79) and (4.6) remain unaltered, that instead of (4.16) one gets the following equation for $\dot{R}$:

$$\dot{R}^2 = c^2 R^{-1}(-kR + \tfrac{1}{3}\Lambda R^3 + \alpha), \tag{6.10}$$

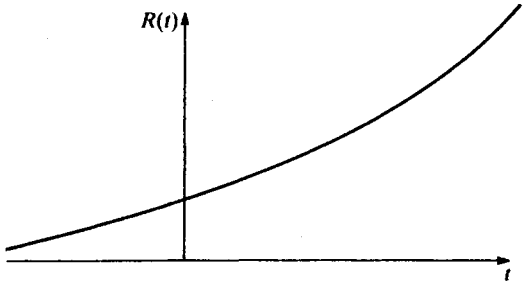


Fig. 6.1. Behaviour of $R(t)$ in the de Sitter model.

where α is a constant given by $\alpha = R_0^3(H_0^2c^{-2} - \frac{1}{3}\Lambda + k/R_0^2)$. The behaviour of the solution depends on the pattern of the zeros and turning points of the cubic on the right hand side of (6.10). There are three particular cases of interest, which are dealt with in the following.

(i) de Sitter model

This arises in the case $k=0$, $\alpha=0$. With the use of (6.8), (6.9) one can show (in the case $p_0=0$), that

$$8\pi G\varepsilon_0 = 3(kc^2/R_0^2 + H_0^2)c^2 - \Lambda c^4. \tag{6.11}$$

Thus if $k=0$ and $\alpha=0$, that is, $H_0^2 = \frac{1}{3}\Lambda c^2$, then the mass-energy density also vanishes, and $R(t)$ is proportional to an exponential:

$$R(t) \propto \exp[(\Lambda/3)^{1/2}tc]. \tag{6.12}$$

One gets a similar form for $R(t)$ in the so-called Steady State Theory of Bondi and Gold (1948) and of Hoyle (1948). However, unlike the de Sitter model, which is empty, in the Steady State Theory there is continuous creation of matter due to the so-called *C*-field.

An interesting property of the metric given by (6.12) is that there is no singularity at a finite time in the past, that is, $R(t)$ does not vanish for any finite value of t (see Fig. 6.1). One can show that this metric has a ten-parameter group of isometries, which is equivalent to 'rotations' in a five-dimensional space with metric whose diagonal elements are $(1, -1, -1, -1, -1)$ and non-diagonal elements zero. This is therefore known as the *de Sitter group*.

(ii) Lemaître model (Lemaître, 1927, 1931)

This model corresponds to the solution of (6.10) with $k=1$ and $\alpha > \alpha_0$, where α_0 is the value of α obtained when Λ has the value in the Einstein static case given by (6.3). From (6.10) we find by differentiation

$$\ddot{R} = +\tfrac{1}{3}c^2\Lambda R - c^2\alpha/2R^2. \tag{6.13}$$

In this model $R(t)$ starts from zero at $t=0$ and increases at first like $t^{2/3}$, that is, initially we can put $R(t)=\xi t^{2/3}$ for some constant ξ. Equation (6.13) then becomes

$$\ddot{R} = \tfrac{1}{3}c^2\Lambda\xi t^{2/3} - (c^2\alpha/2\xi^2)t^{-4/3}. \tag{6.13a}$$

The first term on the right hand side approaches zero and the second term approaches minus infinity as t tends to zero from above. We thus see from (6.13) that at the initial stage $\ddot{R}$ is negative so the expansion is slowing down. The minimum rate of expansion occurs at $R=(3\alpha/2\Lambda)^{1/3}$, when $\ddot{R}=0$, after which the expansion speeds up, ultimately reaching the de Sitter behaviour given by (6.12). An interesting property of this solution is that there is a 'coasting period' near the point at which $\dot{R}$ has its minimum, when the value of $R(t)$ remains almost equal to $(3\alpha/2\Lambda)^{1/3}$. By taking $R(t)$ close to this value, we can write an approximate form of the differential equation (6.13) for $k=1$, as follows:

$$\dot{R}^2/c^2 \simeq -1 + (9\alpha^2\Lambda/4)^{1/3} + \Lambda(R-(3\alpha/2\Lambda)^{1/3})^2, \tag{6.14}$$

which has the following solution:

$$R(t) = (3\alpha/2\Lambda)^{1/3}\{1+[1-(9\alpha^2\Lambda/4)^{-1/3}]^{1/2}\sinh(\Lambda^{1/2}(t-t_{\mathrm{m}})c)\}, \tag{6.15}$$

where t_{m} is the time at which $\dot{R}$ reaches its minimum. By taking α sufficiently close to $2/(3\Lambda)^{1/2}$, one can make the coasting period arbitrarily long. In the latter half of the 1960s there was some evidence that an excess of quasars with red-shifts approximately equal to 2 might exist. This prompted Petrosian, Salpeter and Szekeres (1967) to invoke the Lemaître model, because in this model the parameters can be adjusted so that the 'coasting period' causes an excess of quasars with red-shift 2 or so. However, later the statistical evidence for such an excess disappeared. (See Fig. 6.2 for this model.)

(iii) Eddington–Lemaître model

This is a limiting case of the Lemaître models, which is given this name because it was emphasized by Eddington (1930). In this case $k=1$ and $\alpha=2/(3\Lambda)^{1/2}$, which are the values that obtain in the Einstein static model. This model has an infinitely long 'coasting period'. Thus if $R(0)=0$, then $R(t)$ approaches the Einstein value $(3\alpha/2\Lambda)^{1/3}$ asymptotically from below as t tends to infinity, while if $R(0)=(3\alpha/2\Lambda)^{1/3}$, then $R(t)$ increases monotonically, eventually reaching the de Sitter behaviour as t tends to infinity

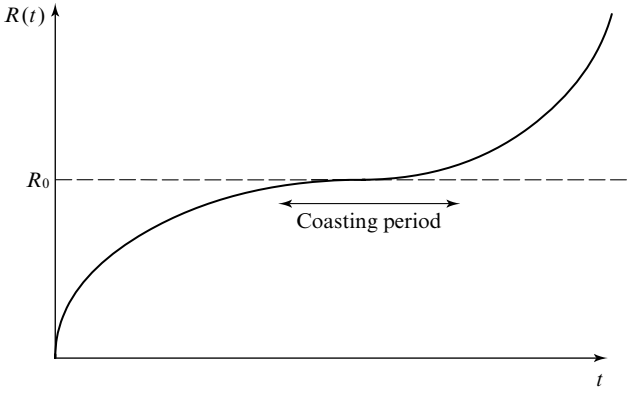


Fig. 6.2. Behaviour of $R(t)$ in the Lemaître models.

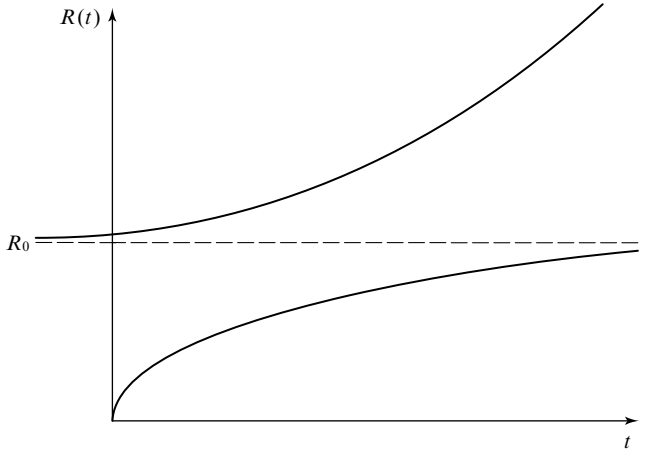


Fig. 6.3. Behaviour of $R(t)$ in the Eddington–Lemaître model.

(see Fig. 6.3). This also shows that the Einstein static model is unstable, at least under perturbations which preserve the Robertson–Walker form of the metric, because when perturbed it will either keep on expanding or approach the Einstein static universe asymptotically.

6.2 Further remarks about the cosmological constant

As is clear from the existence of the Einstein static model, a positive cosmological constant as introduced here represents a repulsive force, so that

the attractive force of the matter is balanced by this repulsive force in the Einstein model. In the dynamic models when the galaxies are very far apart after a period of expansion, the attractive force of the matter becomes weak and eventually the repulsive force due to a positive cosmological constant takes over, and one gets asymptotically the de Sitter behaviour. Correspondingly, with a negative cosmological constant one gets an attractive force in addition to the gravitational attractive force already present.

We saw in the case of the Friedmann models that a model which expands forever corresponds to $k=0, -1$, that is, it has infinite spatial volume (the spatial curvature is zero or negative), whereas a model which eventually collapses has $k=1$, that is, finite spatial volume and positive curvature. This is no longer valid in models with a cosmological constant. We have seen in the case of the Lemaître models (see Fig. 6.2), that although $k=1$, the model expands forever. With a negative cosmological constant it is possible to have $k=0, -1$, and a collapse in the future. This is clear from (6.10), because if Λ is negative eventually the term $\frac{1}{3}\Lambda R^3$ will dominate, so that $R(t)$ cannot be very large, for then $\dot{R}^2$ becomes negative. This will happen regardless of the value of k.

During the mid-seventies there was some evidence that the deceleration parameter q_0 might be negative, that is, the rate of expansion may be increasing. This prompted Gunn and Tinsley (1975) to invoke a Lemaître model with a positive Λ. However, later considerations of evolutionary effects such as that of galactic cannibalism mentioned in the last chapter modified the value of q_0, so that such an 'accelerating' universe no longer seemed necessary. The situation may change again; we will discuss this in the Appendix at the end of the book.

The extent to which a cosmological constant is necessary is uncertain. However, to give cosmological studies generality and scope it seems reasonable to consider (H_0, q_0, Λ) as the three unknown parameters of cosmology which have to be determined from observation. The cosmological constant may turn out to be zero, in which case the actual model will be a pure Friedmann one. However, it may also turn out to be non-zero, for, while there is no compelling reason for having a cosmological constant, there is also not sufficient reason for its absence. Zel'dovich (1968) has suggested that a term may occur due to quantum fluctuations of the vacuum; in this case the cosmological term becomes a part of the energy–momentum tensor. To consider another motivation for having a Λ term, which arises in the work of Hawking (see citation in Islam (1983b)), we have to know about anti-de Sitter space, which occurs in a similar manner to the

de Sitter space considered above except that Λ is now negative. The metric in this case can be written as (A is a constant with dimension of length)

$$ds^2 = c^2\,dt^2 - A^2\cos^2 t[d\chi^2 + \sinh^2\chi(d\theta^2 + \sin^2\theta\,d\phi^2)]. \tag{6.16}$$

This coordinate system covers only a part of the space. For more details on anti-de Sitter space the reader is referred to Hawking and Ellis (1973, p. 131). Hawking considers $N=8$ supergravity theory (see, for example, Freund, 1986) and shows that in this theory a phase transition occurs at a certain critical value of the coupling constant and below this critical value the ground state is an anti-de Sitter space with a negative cosmological constant. Above the critical value there exists a contribution to the Ricci tensor due to vacuum fluctuations which is equivalent to a positive cosmological constant so that the net effect is that the ground state has an 'apparent' cosmological constant which is zero. Other contexts in which the cosmological constant arises will be mentioned later in this book. Other people have given reasons why Λ is small or zero (Coleman, 1988; Banks, 1988; Morris, Thorne and Yurtsever, 1988). Coleman, for example, suggests quantum tunnelling between separate universes (see Chapter 9 and Schwarzschild, 1989).

It was shown by McCrea and Milne (1934) that many of the properties of the Friedmann and Lemaître models can be derived from purely Newtonian considerations if one assumes that the universe is in a dynamic state. The cosmological term is introduced by postulating a force which is proportional to the distance between particles (see the next section). However, the conceptual basis of this formulation is not sound partly because it does not incorporate the special theory of relativity.

It is clear from the above discussion that it is important to have limits on the cosmological constant. This we will consider in the next section. For a selection of other works on Lemaître models, we refer to Petrosian and Salpeter (1968), Kardashev (1967), Brecher and Silk (1969), Tinsley (1977), Raychaudhuri (1979) and Bondi (1961) (the last three contain reviews).

6.3 Limits on the cosmological constant

From (6.7) and (6.9) we get the following relation:

$$q_0 = \tfrac{1}{2}\Omega_0 - c^2\Lambda/3H_0^2. \tag{6.17}$$

This is the equation which replaces (4.14), the latter being valid for Friedmann models. From (6.17) we get

$$|q_0 - \tfrac{1}{2}\Omega_0| = |c^2\Lambda/3H_0^2|. \tag{6.18}$$

Although the observational values of q_0 and Ω_0 are uncertain, one can reasonably safely say that q_0 lies between -5 and $+5$, and that Ω_0 lies between 0 and 4. The left hand side of (5.18) can then have the maximum value of 7, so that we get

$$|\Lambda| = 21 H_0^2/c^2. \tag{6.19}$$

By setting a limit of 100 km s^{-1} Mpc^{-1} on H_0, (6.19) leads to a limit of approximately 10^{-54} cm^{-2} on the absolute value of Λ (this limit is mentioned by Hawking; see citation in Islam (1983b)).

The above limit comes from cosmological considerations. It is of some interest to see if local considerations can give anything like the same limits. Such a local limit can be obtained by considering the effect of a Λ term on the perihelion shift of Mercury (Islam, 1983b). The Schwarzschild metric is modified as follows by the Λ term (here r has dimension of length):

$$\begin{aligned} ds^2 &= c^2(1 - 2m/r - \tfrac{1}{3}\Lambda r^2)dt^2 \\ &\quad - (1 - 2m/r - \tfrac{1}{3}\Lambda r^2)^{-1}dr^2 - r^2(d\theta^2 + \sin^2\theta\, d\phi^2), \end{aligned} \tag{6.20}$$

where m is the mass of the Sun, multiplied by G/c^2. It is well known that the usual Schwarzschild solution implies a perihelion shift of Mercury of about 43″ per century. This shift is known with an accuracy of about half a per cent. Using this fact one can show that Λ must satisfy the following inequality (see Islam (1983b) for more details):

$$|\Lambda| < 10^{-42}\ \text{cm}^{-2}. \tag{6.21}$$

Thus the limit from local considerations is much worse than that derived from cosmology, as expected. One can improve on (6.21) by considering local systems of bigger dimensions, such as the fact that a galaxy is a bound system (Islam and Munshi, 1990; Munshi, 1999). For this we consider a typical galaxy such as ours with 10^{11} stars of solar mass, that is, of mass 2×10^{33} g. The matter contained in a disc of diameter 80 thousand light years and thickness 6 thousand light years we imagine to fill a sphere of uniform density with the same average density. The equivalent sphere has a radius of about 19 thousand light years.

Let $\mathbf{r}$ be the position vector of a point with respect to the centre of the spherical galaxy, then we assume the force on a unit mass to be given by (ρ is the density):

$$\mathbf{F} = -\tfrac{4}{3}\pi\rho G\mathbf{r} + \tfrac{1}{3}\Lambda c^2\mathbf{r}. \tag{6.22}$$

Here the Λ term is the Newtonian form of the cosmological term. As before, a positive Λ implies a repulsive force. The first term on the right

hand side of (6.22) represents the usual gravitational force. The galaxy ceases to be a bound system if the right hand side of (6.22) gives an outward force. The condition for this is

$$\Lambda < 4\pi\rho G/c^2. \tag{6.23}$$

For the dimensions given above we find $4\pi\rho \approx 1.14 \times 10^{-22}$ g cm^{-3}, so that (6.23) gives a limit of approximately 10^{-48} cm^{-2} for Λ. This could also be considered as a limit on the absolute value of Λ, for even if Λ is negative, if its absolute value violated this limit, the effect on the binding of the galaxy would be noticeable. One has to augment this analysis with a general relativistic one by considering the Schwarzschild interior solution, and its modification due to the Λ term. For this and other details we refer to Islam and Munshi (1990) and Munshi (1999).

6.4 Some recent developments regarding the cosmological constant and related matters

6.4.1 Introduction

In the preceding sections we have provided some basic information regarding the cosmological constant, including some historical aspects. In this section we consider some more recent developments which have both theoretical and observational aspects; the latter can be considered as extensions of observational cosmology discussed in the last chapter. For convenience we may repeat some earlier remarks. There are some uncertainties, as usual, both theoretically and observationally, but we will attempt to present a balanced picture and try to convey the 'flavour' of the current research. We will rely largely on the reviews by Carroll, Press and Turner (1992), Weinberg (1989), Bagla, Padmanabhan and Narlikar (1996), and Viana and Liddle (1996). We also present an exact solution with the cosmological constant.

For simplicity we restrict to the zero pressure case: $p = 0$. In this case (see (4.15), which is valid for $\Lambda \neq 0$, as mentioned earlier – see also (4.6)), we get

$$\varepsilon/\varepsilon_0 = (R_0/R)^3 = \rho_{\mathrm{M}}/\rho_{\mathrm{M0}}, \tag{6.24}$$

where we have introduced the mass density ρ_{M} related to the mass-energy density ε as follows. In the case of non-zero pressure p the mass-energy density consists of the rest mass of the particles constituting the matter, the energy density of any radiation present, and the energy arising from the random motion of the particles that occurs when the pressure is non-zero. In the case of zero pressure, that is, dust, assuming radiation to be

absent, the energy density is simply that given by the mass density, and it may be more appropriate to use ρ_{M} with $\varepsilon = \rho_{\mathrm{M}} c^2$, and ρ_{M0} as the present value of ρ_{M}: $\rho_{\mathrm{M0}} = \varepsilon_0 / c^2$.

Consider (6.6) and (6.11) and write the latter as follows:

$$(8\pi G\varepsilon_0)/(3c^2H_0^2) = kc^2/(H_0^2R_0^2) + 1 - \Lambda c^2/(3H_0^2). \tag{6.25}$$

The left hand side is the present value of the density parameter Ω_0 (see remarks following (4.9) and (6.6)), which can also be written as follows: $\Omega_0 = (8\pi G\rho_{\mathrm{M0}})/(3H_0^2)$. With the following definitions:

$$\Omega_\Lambda = (\Lambda c^2)/(3H_0^2);\ \Omega_{\mathrm{k}} = -kc^2/(H_0^2R_0^2), \tag{6.26}$$

Equation (6.25) can be written as follows:

$$\Omega_0 + \Omega_{\mathrm{k}} + \Omega_\Lambda = 1. \tag{6.27}$$

Introduce the following dimensionless forms of R and t, given respectively by a and τ:

$$a(\tau) = R(t)/R_0;\ \tau = H_0 t. \tag{6.28}$$

We proceed to derive a first order differential equation for $a(\tau)$. We have

$$\mathrm{d}a/\mathrm{d}\tau = (\mathrm{d}R/\mathrm{d}t)(\mathrm{d}a/\mathrm{d}R)(\mathrm{d}t/\mathrm{d}\tau) = \dot{R}/(R_0H_0). \tag{6.29}$$

Next we re-write (6.10), inserting the value of α given after the equation:

$$\dot{R}^2 = -c^2k + (\tfrac{1}{3})c^2\Lambda R^2 + R_0^3H_0^2R^{-1} - (\tfrac{1}{3})c^2\Lambda R_0^3R^{-1} + c^2kR_0R^{-1}. \tag{6.30}$$

We divide this equation by $R_0^2H_0^2$ and re-arrange terms to get the following equation:

$$\begin{aligned}\dot{R}^2/(R_0^2H_0^2) = {} & (-c^2k/R_0^2H_0^2)(1 - R_0R^{-1}) \\ & + (c^2\Lambda/3H_0^2)(R^2/R_0^2 - R_0R^{-1}) + R_0R^{-1}.\end{aligned} \tag{6.31}$$

This equation can be expressed readily in terms of a, $\mathrm{d}a/\mathrm{d}\tau$, Ω_k and Ω_Λ, with the use of (6.26), (6.28) and (6.29). Eliminating Ω_{k} from the resulting equation with the use of (6.27), we get

$$(\mathrm{d}a/\mathrm{d}\tau)^2 = 1 - \Omega_0(1 - a^{-1}) + \Omega_\Lambda(a^2 - 1). \tag{6.32}$$

This can be written as

$$(\mathrm{d}a/\mathrm{d}\tau)^2 = a^{-1}[\Omega_\Lambda a^3 + (1 - \Omega_0 - \Omega_\Lambda)a + \Omega_0], \tag{6.33}$$

a form that leads to the integral solution (6.34).

In the next subsection we present an exact solution which might not be physically important, but may provide a useful exercise for the reader.

6.4.2 An exact solution with cosmological constant

Before proceeding with the observational aspect, we consider (6.32) further. The case $\Lambda = 0$ gives the Friedmann models discussed in the last chapter, while some approximate and limiting solutions for $\Lambda \neq 0$ were considered earlier in this chapter. A general solution of (6.32) in the form of an integral can be expressed as follows:

$$(\tau - \tau_i) = \pm \int_{a_i}^{a} a^{\frac{1}{2}} \mathrm{d}a / [\Omega_\Lambda a^3 + (1 - \Omega_0 - \Omega_\Lambda)a + \Omega_0]^{\frac{1}{2}}, \tag{6.34}$$

where τ_i, a_i are some initial values of τ, a, e.g., those at the big bang. The right hand side is in general an elliptic integral which cannot be expressed in terms of elementary functions. However, there is a fortuitous combination of values of Ω_0, Ω_Λ for which the integral can be evaluated. Although these values might be unrealistic, an explicit evaluation may give some idea of the form of the function (for some parameter values) which may be of interest in some contexts. The solution consists of two parts, for positive and negative Λ; the former is related to the approximate solution considered in Section 6.1 (see (6.15) and Fig. 6.2)). In fact for positive Λ the exact solution describes the two curves in Fig. 6.3, depending on the boundary condition at the origin. We write the cubic in (6.34) as follows:

$$\Omega_\Lambda a^3 + (1 - \Omega_0 - \Omega_\Lambda)a + \Omega_0 = \Omega_\Lambda (a - \alpha')^2 (a + \beta), \tag{6.35}$$

where α' and β are constants. It is readily verified that in this case we must have

$$\alpha' = (\Omega_0 / 2\Omega_\Lambda)^{1/3} = (\tfrac{1}{2})\beta, \tag{6.36a}$$

and Ω_0, Ω_Λ must satisfy a relation which leads to the value of Λ as follows, with $k = 1$, assuming α' to be positive.

$$\begin{aligned} \Lambda &= (4c^4/9R_0^6H_0^4)\,(\rho_c/\rho_{\mathrm{M}0})^2;\ 1 - \Omega_0 - \Omega_\Lambda = \Omega_k \\ &= -k/R_0^2H_0^2 = -3(\Omega_0^2\Omega_\Lambda/4)^{1/3}, \end{aligned} \tag{6.36b}$$

where $\rho_c = (3H_0^2/8\pi G)$, the critical value of ρ. The relation (6.34) then reduces to the following; the positive sign is more appropriate:

$$\Omega_\Lambda^{-\frac{1}{2}} \int a^{\frac{1}{2}} \mathrm{d}a / [(a - \alpha')\,(a + 2\alpha')^{\frac{1}{2}}] = (\tau - \tau_i). \tag{6.37}$$

The substitution

$$a = 2\alpha' \sinh^2 \psi, \tag{6.38a}$$

transforms the integral on the left hand side to the following one:

$$\int(4\sinh^2\psi\,\mathrm{d}\psi)/(2\sinh^2\psi-1)=2\int[\mathrm{d}\psi+\mathrm{d}\psi/(2\sinh^2\psi-1)]. \quad (6.38b)$$

The second integral on the right hand side can be written as follows and evaluated through the substitution $\mathrm{e}^{2\psi}=\xi$:

$$\int 2\mathrm{e}^{2\psi}\mathrm{d}\psi/(\mathrm{e}^{4\psi}-4\mathrm{e}^{2\psi}+1)=\int \mathrm{d}\xi/[(\xi-2)^2-3]$$
$$=(1/(2\sqrt{3}))\log[(\xi-2-\sqrt{3})/(\xi-2+\sqrt{3})]+\text{constant}. \quad (6.39)$$

In terms of ψ given by (6.38a), the equation (6.37) can thus be written as:

$$(\tau-\tau_i)=\Omega_\Lambda^{-\frac{1}{2}}\log\{\mathrm{e}^{2\psi}[A(\sqrt{3}+2-\mathrm{e}^{2\psi})/(\mathrm{e}^{2\psi}-2+\sqrt{3})]^{1/\sqrt{3}}\}, \quad (6.40)$$

where $A=(\sqrt{3}-1)/(\sqrt{3}+1)$, so chosen that τ_i represents the moment of the big bang when $a=0$. If $\tau_i=0$, then $a=0$ when $\tau=0$ (see (6.38a)). This choice leads to the lower branch in Fig. 6.3. With the use of (6.38a) ψ can be expressed in terms of a as follows:

$$\mathrm{e}^{2\psi}=[1+a/\alpha'\pm(a^2/\alpha'^2+2a/\alpha')^{\frac{1}{2}}]. \quad (6.41)$$

This can be substituted in (6.40) to express τ in terms of a. Let us consider behaviour near $\tau=0$ (setting $\tau_i=0$) and $a=0$. To this end we expand the right hand side of (6.41) in powers of a, as follows:

$$\mathrm{e}^{2\psi}=[1\pm(2a/\alpha')^{\frac{1}{2}}+a/\alpha'\pm(1/2\sqrt{3})\,(a/\alpha')^{3/2}+\cdots]. \quad (6.42)$$

It can be verified that terms of order $a^{\frac{1}{2}}$ cancel when the expression is substituted in (6.40). The next order term leads to a linear behaviour for a on the right hand side of (6.40), so that this would lead to a solution in which the scale factor behaves like τ (or t) near the origin, i.e., near the big bang. However, such a behaviour would imply that $(\mathrm{d}a/\mathrm{d}\tau)$ is finite at $\tau=0$, which is inconsistent with (6.32), from which it is clear that $(\mathrm{d}a/\mathrm{d}\tau)$ is infinite at $\tau=0$. Indeed, the term linear in a on the right hand side of (6.40) also vanishes, as can be verified. The next term is of order $a^{3/2}$, and this leads to the behaviour $a\sim\tau^{2/3}$, which is consistent with (4.45) and the fact, evident from (6.32), that the Λ term does not affect behaviour near $\tau=0$. We will come back to this solution after considering the case where α' in (6.35) is negative.

Let α' in (6.35) be negative and set $-\alpha'=\hat{\alpha}>0$. From (6.36a) we see that Ω_Λ is negative and accordingly we write (6.32) using (6.35) with $\hat{\alpha}=-\alpha'$, as follows:

$$(\mathrm{d}a/\mathrm{d}\tau)^2=-\Omega_\Lambda a^{-1}(a+\hat{\alpha})^2(2\hat{\alpha}-a). \quad (6.43)$$

The second part of (6.36b), which is still valid, implies $k=-1$. This leads to the following integral (we choose the positive sign in (6.45)):

$$\tau-\tau_i=\pm(-\Omega_\Lambda)^{-\frac{1}{2}}\int_{a_i}^{a} a^{\frac{1}{2}}\mathrm{d}a/[(a+\hat{\alpha})\,(2\hat{\alpha}-a)^{\frac{1}{2}}], \tag{6.44}$$

with again τ_i, a_i some initial values of τ, a, which can both be taken as zero, as we do in the following. The substitution $a=2\hat{\alpha}\,\sin^2\theta$ yields

$$\tau=(-\Omega_\Lambda)^{-\frac{1}{2}}\int\{2-2/(1+2\,\sin^2\theta)\}\,\mathrm{d}\theta. \tag{6.45}$$

The substitution $\eta=\tan(\theta/2)$ transforms the second integral as follows:

$$\begin{aligned}&\int \mathrm{d}\theta/(1+2\,\sin^2\theta)=\int 2(1+\eta^2)\mathrm{d}\eta/(1+10\eta^2+\eta^4)\\&=\int \mathrm{d}\eta\{(1+(2/3)^{\frac{1}{2}})/(1+(5+2\sqrt{6})\eta^2)+(1-(2/3)^{\frac{1}{2}})/(1+(5-2\sqrt{6})\eta^2)\}.\end{aligned} \tag{6.46}$$

The integrals are now standard; performing the integration and substituting for η, we get the following implicit relation between τ and a (with $\theta=\sin^{-1}(a/2\hat{\alpha})$)

$$\begin{aligned}\tau=(-\Omega_\Lambda)^{-\frac{1}{2}}\Bigg\{&2\theta-\frac{2(1+(2/3)^{\frac{1}{2}})}{(5+2\sqrt{6})^{\frac{1}{2}}}\tan^{-1}\left[(5+2\sqrt{6})^{\frac{1}{2}}\tan\frac{\theta}{2}\right]\\&-\frac{2(1-(2/3)^{\frac{1}{2}})}{(5-2\sqrt{6})^{\frac{1}{2}}}\tan^{-1}\left[(5-2\sqrt{6})^{\frac{1}{2}}\tan\frac{\theta}{2}\right]\Bigg\}.\end{aligned} \tag{6.47}$$

The corresponding expression for the case $\Lambda>0$ ($\alpha'>0$) can be written down by first expressing (6.40) as follows (with $\tau_i=0$)

$$\tau=(-\Omega_\Lambda)^{-\frac{1}{2}}\{2\psi+(1/\sqrt{3})\log(A(\sqrt{3}+2-\mathrm{e}^{2\psi})/(\mathrm{e}^{2\psi}-2+\sqrt{3}))\}, \tag{6.48}$$

with $A=(\sqrt{3}-1)/(\sqrt{3}+1)$, and then substituting for ψ from (6.41), as indicated earlier. In (6.47) and (6.48), it can be verified that $\tau=0$ implies $a=0$.

We consider briefly some properties of the solution. From (6.32) and (6.35) we see, for $\alpha'>0$, that

$$(\mathrm{d}a/\mathrm{d}\tau)^2=\Omega_\Lambda a^{-1}(a-\alpha')^2(a+2\alpha'). \tag{6.49}$$

Taking derivatives with respect to τ, it is clear, because of the $(a-\alpha')^2$ factor, that both $(\mathrm{d}a/\mathrm{d}\tau)$ and $(\mathrm{d}^2a/\mathrm{d}\tau^2)$ vanish at $a=\alpha'$. This is therefore a point of inflexion for the curve (6.49), or the latter is asymptotic to the line $a=\alpha'$. In fact the second situation is the correct one, and leads to the behaviour displayed in Fig. 6.3.

In the case $-\alpha'=\hat{\alpha}>0$ (6.32) reduces to (6.43). In this case we have $(\mathrm{d}a/\mathrm{d}\tau)=0$ for $a=2\hat{\alpha}$, but, as is readily verified, $(\mathrm{d}^2a/\mathrm{d}\tau^2)$ does not vanish

for any value of a. We then get a 'closed' model akin to the case $k=1$ in Fig. 4.2; the maximum value of a is $2\hat{\alpha}$.

The considerations of this subsection and extensions of these may provide 'comprehension exercises' for the reader to get to know better the graphical and analytic structure of the Friedmann and Lemaître models.

6.4.3 Restriction of parameter space

As indicated, several sets of authors have studied and reviewed the observational and related theoretical situation to try and restrict the 'space' of observational parameters, to see which set of models has more validity than others. As a representative such review, we shall follow that of Bagla, Padmanabhan and Narlikar (1996). We choose this review partly because it is concise and clear, and not necessarily because we regard it as the most accurate one. This can be considered as one of several 'platforms' from which to assess other reviews and the overall situation. Because of the observational and theoretical uncertainties, as well as the natural tendency to emphasize certain aspects, there is usually a subjective element in the reviews. Later we will mention some more recent reviews.

Bagla *et al.* (1996) start by mentioning a review by Gunn and Tinsley (cited earlier, on p. 99) carried out in 1975 in which they conclude: 'new data on the Hubble diagram, combined with constraints on the density of the universe and the ages of galaxies, suggest that the most plausible cosmological models have a positive cosmological constant, are closed, too dense to make deuterium in the big bang, and will expand for ever . . .'. Because of various developments in the intervening period, the reviewers feel that the time is ripe to 'take fresh stock of the cosmological situation today'. Indeed, certain new aspects have come to the fore which were not seriously considered a decade or two ago. One of these is the abundance of rich clusters of galaxies, some of which contain as much as 10^{15} solar masses. As Viana and Liddle (1996) say, 'One of the most important constraints that a model of large-scale structure must pass is the ability to generate the correct number density of clusters. This is a crucial constraint . . .' (see also Liddle *et al.*, 1996). Other ingredients that go into the examination of a model are: consistency with the ages of the oldest objects in the universe, namely, globular clusters, whose age has been estimated to be 15.8 ± 2.1 billion years (Bolte and Hogan, 1995; Bolte, 1994), fraction of mass contributed by baryons in rich clusters, abundance of high red-shift objects in radio galaxies and so-called damped Lyman alpha systems (DLAS) (Lanzetta, Wolfe and Turnshek, 1995), and, of course, more

recent measurements of the Hubble constant H_0. For a 'local' value of H_0 (that is, from studies of relatively nearby objects) Freedman *et al*. (1994) get $H_0 = 80 \pm 17$, while a 'global' value (from distant objects) of $H_0 = 65 \pm 25$ was obtained by Birkinshaw and Hughes (1994) (compare Sandage and Tamman, 1975 (see (3.63)); see also Saha *et al*., 1995). We will return to some of these points later.

Bagla *et al*. analyse in detail essentially two models: (i) the first one satisfies $\Omega_0 + \Omega_\Lambda = 1$, $k = 0$, and (ii) the second satisfies $\Omega_0 < 1$, $\Omega_\Lambda = 0$, $k = -1$ (see (6.33)). The results are set out in Fig. 6.4, which is a simplified and modified version of their Fig. 3 (Bagla *et al*., 1996). The Hubble constant here is given in units of 100, that is, $h = H_0/100$ km s^{-1} Mpc^{-1}. The upper and lower boxes correspond to the cases (i) and (ii) cited above.

As mentioned, this review is meant to take stock of the situation in observational cosmology at the time of writing (1996). As will be clear from Chapter 9, a strong requirement for inflation is that the density parameter should equal unity. In recent times this requirement has been somewhat modified to include the cosmological constant, so that $\Omega_{\text{tot}} = \Omega_0 + \Omega_\Lambda$ should equal unity (see 6.33); this is a sort of 'generalized' flatness condition. This is incorporated in model (i). Model (ii) does not necessarily conform to the inflation condition and has zero cosmological constant and negative curvature. The two models therefore are of somewhat different nature and so representative of a wide spectrum. To recapitulate, the observational constraints being used here are: the Hubble constant, the deceleration parameter, ages of globular clusters, abundance of primordial deuterium and of rich clusters, baryon content of galaxy clusters and abundance of high red-shift objects. The authors find that the available parameter space is rather limited. This casts doubt on the error bars of the measurements, or requires fine tuning of the theoretical models.

We now discuss in some detail the ingredients that go into Fig. 6.4. The topic of the cosmic background radiation (CBR), mentioned in Chapter 1, will arise; this will be dealt with in detail in Chapter 8. Here we refer to it in broad terms. Firstly, the age of globular clusters, which are known to be amongst the oldest objects in the universe, set a lower limit to the age of the universe. The ages of the stars in the globular clusters can be calculated from their mass, from the observed metallicity, and the point at which they leave the well-known, so-called 'main sequence' in the Herzsprung–Russell (HR) diagram. Nuclear reactions in a star result in heavier nuclei, until one gets to iron, which has the most stable nucleus. Metallicity thus gives indication of the stage of nuclear burning. Bolte and Hogan (1995) estimate the ages of stars in M92 to lie in 15.8 ± 2.1 Gyr (billion years). Given

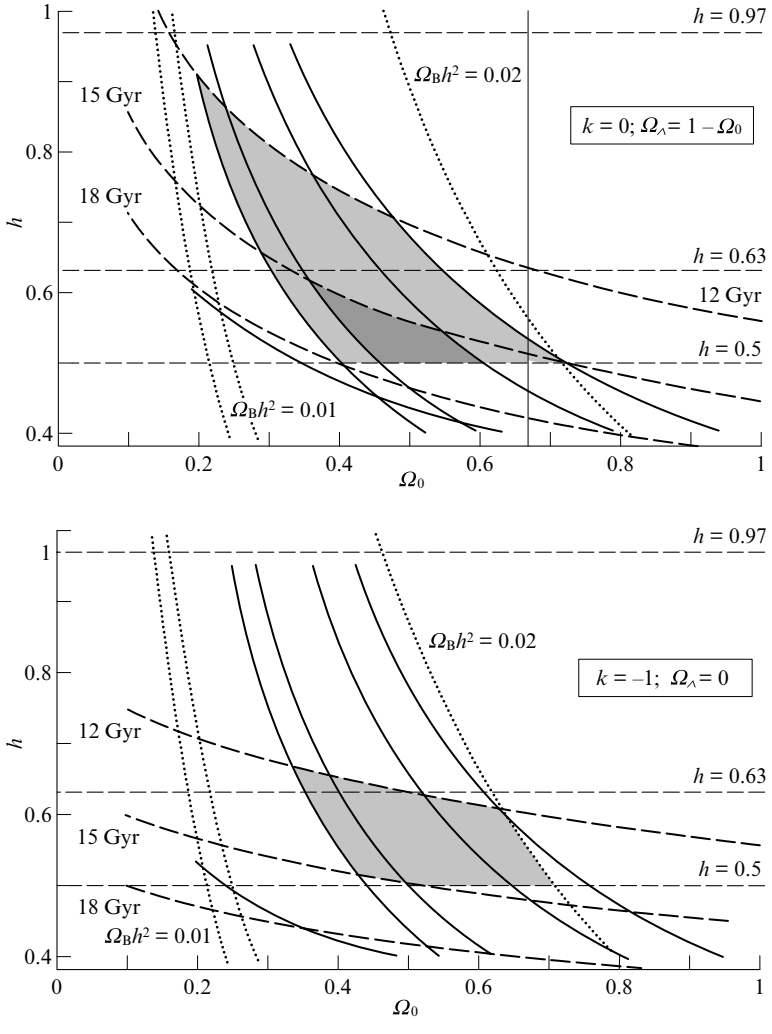


Fig. 6.4. This diagram displays the restrictions on the (h, Ω_0) plane imposed by various observations, as worked out by Bagla, Padmanabhan and Narlikar (1996). The upper part refers to a model with $k=0$, $\Omega_0+\Omega_\Lambda=1$, while the lower one describes a model with $k=-1$, $\Omega_\Lambda=0$. The text (Section 6.4.3) explains the various curves.

the values of Ω_0, Ω_Λ, and H_0, the age t_0 of the universe can be calculated; these are indicated by dashed lines in the figure for $t_0 = 12$, 15 and 18 Gyr; the allowed region in the diagram lies below the corresponding curve. The upper diagram displays these curves for models with $k = 0$, while the lower one is for models with $k = -1$, $\Omega_\Lambda = 0$.

The measurement of distance to M100, which is a galaxy in the Virgo cluster, using the Cepheid luminosity relation and the Hubble Space Telescope (HST), gives $h = 0.80 + 0.17$ (Freedman *et al.*, 1994). This may be regarded as a 'local' value which may be different from the global value. Turner *et al.* (1991) and Nakamura and Suto (1995) have estimated the probability distribution of a local value of H_0; apparently values smaller than $h = 0.5$ can be ruled out at 94% confidence. Birkinshaw and Hughes (1994) find a 'global' value of $h = 0.65 \pm 0.25$, for Abell 2218, with methods like the Sunyaev–Zel'dovich effect; the latter is caused by the scattering of CBR photons by electrons in hot plasmas that exist in some regions. This produces an observable effect (Sunyaev and Zel'dovich, 1980). Sandage and Tamman, as mentioned earlier, get values of h in the range 0.5–0.6, using various methods. In the figure horizontal dotted lines give limits on values of h.

One of the important large scale properties of the universe is the occurrence of rich clusters of galaxies; these can be identified from X-ray observations if one assumes the central temperature to exceed 7 keV. The abundance of rich clusters can be worked out theoretically. The calculations have some model dependence, hence the possibility of confronting models with observations. The mass of the clusters is estimated, for example, by assuming virial equilibrium and using the velocity dispersion of galaxies, gravitational lensing, etc. The cluster abundance can be calculated analytically by the Press–Schechter theory (Press and Schechter, 1974; Bond *et al.*, 1991; Bond, 1992, 1995). This theory gives the fraction of material contained in gravitationally bound systems larger than some mass M, in terms of the fraction of space where the linearly evolved (suitably smoothed out) density field exceeds some threshold (see Eq. (13) of Viana and Liddle, 1996).

An alternative way to compare observation with theory is to convert the number density of clusters into amplitude of density fluctuations, scaled in a suitable manner. One assumes a power law for the root-mean-square density perturbations (White *et al.*, 1993). The constraints arising from these considerations are represented by a pair of thick unbroken lines. The thin pair of outer lines indicate uncertainties arising from normalization of certain COBE (Cosmic Background Explorer, about which more later) data.

The Coma cluster is a rich cluster of galaxies which has been studied in great detail. This cluster can be considered to be a prototype. From these studies it emerges that the fraction of mass contributed by baryons is given by

$$(\text{Mass in baryons/Total mass}) = (\Omega_B/\Omega_0) = 0.009 + 0.050\text{h}^{-3/2}, \tag{6.50}$$

where the right hand side is uncertain by 25%. As we shall see in Chapter 8, light nuclei formed in the early universe; the abundance of different elements is a function of $\Omega_B h^2$. Thus the value of Ω_B obtained from primordial nucleosynthesis can be used in conjunction with (6.50) to put additional constraints on Ω_0. The lowest and highest such bounds are represented by thick dot-dashed lines; the permitted region lies to the left in each case. More recent observations of deuterium in a high red-shift system (Saha *et al.*, 1995) imply values of Ω_B smaller than previous ones. The resulting constraint is represented by the thin dot-dashed line in Fig. 6.4.

The occurrence of high red-shift objects such as radio galaxies and damped Lyman alpha systems (DLAS) imply that the amplitude of density perturbations at $z=2$ is of order unity at $M \cong 10^{11}\text{M}_\odot$. This circumstance places a lower bound in the (h, Ω_0) diagram, represented by the unbroken line at the lower left hand corner. For flat models (at the top) this line runs alongside the line of constant age (18 Gyr) and so provides an upper bound for the age of the universe. For DLAS and related matters, we refer to Subramanian and Padmanabhan (1993).

Lastly, for the present, the deceleration parameter, about which there is considerable uncertainty, as mentioned earlier, is represented here by the vertical line in the top diagram. The allowed region lies to the right.

As indicated earlier, the present review can be taken as a platform. As the authors themselves say, any changes in the observations or, to some extent, in the theory, can be taken care of, within reason, by suitably modifying and scaling.

Much of the material of this subsection has been considered in detail in interesting papers by Viana and Liddle (1996) and by Liddle *et al.* (1996). (See also Liddle and Lyth, 2000.)

Some recent observations of supernovae in distant galaxies indicate the existence of a positive cosmological constant and, consequently, a possible accelerating universe (Perlmutter *et al.*, 1998). This will be discussed briefly in the Appendix.

7
Singularities in cosmology

7.1 Introduction

In Chapter 4 we saw that all the Friedmann models have singularities in the finite past, that is, at a finite time in the past, which we have called $t = 0$; the scale factor $R(t)$ goes to zero and correspondingly some physical variables, such as the energy density, go to infinity. Only exceptionally, such as in the de Sitter or the steady state models (see Fig. 6.1), is there no singularity in the finite past. But these latter models have some unphysical or unorthodox feature, such as the continuous creation of matter, which is not generally acceptable. The presence of singularities in the universe, where physical variables such as the mass-energy density or the pressure or the strength of the gravitational field go to infinity seems doubtful to many people, who therefore feel uneasy about this kind of prediction of the equations of general relativity. This was partly the motivation with which Einstein searched for a 'unified field theory'. In this connection he says (1950):

> The theory is based on a separation of the concepts of the gravitational field and matter. While this may be a valid approximation for weak fields, it may presumably be quite inadequate for very high densities of matter. One may not therefore assume the validity of the equations for very high densities and it is just possible that in a unified theory there would be no such singularity.

There was at one time the feeling that the singularities in the Friedmann models arise because of the highly symmetric and idealized form of the metric, and that, for example, if the metric were not spherically symmetric, the matter coming from different directions might 'miss' each other and not gather at the centre of symmetry, as it does in the (spherically symmetric) Friedmann models. However, it was shown by Hawking and Penrose

(1970) that spherical symmetry is not essential for the existence of a singularity. We shall consider this work later.

There are in the main two possible approaches for dealing with the problem of singularities. Firstly, one can try to relax the symmetry conditions inherent in Robertson–Walker metrics and try to determine what the field equations predict in these more general cases. Secondly, one can try to derive some general results about singularities by using reasonable assumptions, say about the energy–momentum tensor, without considering the field equations in detail. The Penrose–Hawking results fall in the latter category. As regards the former approach, the simplest relaxation of the symmetries of the Robertson–Walker metrics (which are homogeneous and isotropic) is to drop the requirement of isotropy and consider metrics which are only homogeneous. A simple example of such a metric was given in (3.15). We shall consider such metrics in some detail in the next section, partly with a view to explaining another approach to the question of singularities, pioneered by Lifshitz and Khalatnikov (1963). There is an extensive literature on singularities and cosmological solutions, incorporating both the approaches mentioned above. This chapter is meant to be only a brief introduction to this work. For more detailed reviews the reader is referred to Hawking and Ellis (1973), Ryan and Shepley (1975), Landau and Lifshitz (1975), MacCallum (1973), Raychaudhuri (1979) and Clarke (1993).

7.2 Homogeneous cosmologies

In this section we shall derive the metric and field equations for homogeneous (but not isotropic) cosmologies. We shall give the bare essentials here. For more details the reader can consult Landau and Lifshitz (1975, p. 381).

In Section 3.1 we defined a homogeneous space. To continue that discussion, consider the spatial part of the metric (3.1), as follows:

$$\mathrm{d}l^2 = h_{ij}(t, x^1, x^2, x^3)\mathrm{d}x^i\,\mathrm{d}x^j, \tag{7.1}$$

where as usual the indices i and j are to be summed over values 1, 2, 3. A metric is homogeneous if after a transformation of the spatial coordinates x^1, x^2, x^3 to new coordinates x'^1, x'^2, x'^3 the metric (7.1) transforms to the following one:

$$\mathrm{d}l^2 = h_{ij}(t, x'^1, x'^2, x'^3)\mathrm{d}x'^i\,\mathrm{d}x'^j, \tag{7.2}$$

with the same functional dependence as before of the h_{ij} on the new spatial coordinates. Further, this set of transformations must be able to carry any

point to any other point. We saw an explicit example of such a transformation in a simple case in (3.15). One way to characterize the invariance of the metric under spatial transformations is to consider a set of three differential forms $e^{(a)}_m \, dx^m$ (with $a = 1, 2, 3$) which are invariant under these transformations, as follows:

$$e^{(a)}_m(x) \, dx^m = e^{(a)}_m(x') \, dx'^m, \tag{7.3}$$

where we have written x for x^1, x^2, x^3, etc., in the arguments. With the use of these forms a metric invariant under spatial transformations can be constructed as follows (the η_{ab} are six functions of t):

$$dl^2 = \eta_{ab}(e^{(a)}_m \, dx^m)(e^{(b)}_n \, dx^n), \tag{7.4}$$

that is, the three-dimensional metric tensor h_{ij} of (7.2) is given as follows:

$$h_{ij} = \eta_{ab} e^{(a)}_i e^{(b)}_j. \tag{7.5}$$

Note that in (7.3) the $e^{(b)}_m$ on the two sides of the equation are respectively the same functions of the old and new coordinates. We introduce the reciprocal triplet of vectors $e^m_{(a)}$ by the following relations:

$$e^m_{(a)} e^{(b)}_m = \delta^b_a, \; e^m_{(a)} e^{(a)}_n = \delta^m_n. \tag{7.6}$$

It can be shown after some manipulations (see Landau and Lifshitz, 1975, pp. 382–3), that (7.3) leads to the following equation for the reciprocal triplet $e^m_{(a)}$:

$$e^m_{(a)} \frac{\partial e^n_{(b)}}{\partial x^m} - e^m_{(b)} \frac{\partial e^n_{(a)}}{\partial x^m} = C^c_{ab} e^n_{(c)}, \tag{7.7}$$

where the C^c_{ab} are constants satisyfing $C^c_{ab} = -C^c_{ba}$. These are the so-called structure constants of the groups of transformations. If we denote by X_a the following linear differential operator:

$$X_a = e^m_{(a)} \frac{\partial}{\partial x^m}, \tag{7.8}$$

then (7.7) can be written as follows:

$$[X_a, X_b] \equiv X_a X_b - X_b X_a = C^c_{ab} X_c. \tag{7.9}$$

One can now use the Jacobi identity given by

$$[[X_a, X_b], X_c] + [[X_b, X_c], X_a] + [[X_c, X_a], X_b] = 0, \tag{7.10}$$

to derive the following relation for the structure constants:

$$C^e_{ab} C^d_{ec} + C^e_{bc} C^d_{ea} + C^e_{ca} C^d_{eb} = 0. \tag{7.11}$$

The different types of homogeneous spaces correspond to the different inequivalent solutions of (7.11) satisfying the antisymmetry condition $C^c_{ab} = -C^c_{ba}$. Some solutions are equivalent to each other, reflecting the fact that the $e^m_{(a)}$ can still be subjected to a linear transformation with constant coefficients so that the operators X_a are not unique.

There are nine different types of homogeneous spaces that arise from the different inequivalent solutions of (7.11) with the required antisymmetry condition. These are known as the Bianchi types, types I–IX. The Einstein equations for these spaces can be reduced to a system of ordinary differential equations for the $\eta_{ab}(t)$, without the necessity of working out the frame vectors $e^{(a)}_m$, etc. We will consider an application of these results in Section 7.7.

7.3 Some results of general relativistic hydrodynamics

Before considering the results of Penrose and Hawking it is useful to have some idea of relativistic hydrodynamics. The fundamental quantity here is the four-velocity vector u^μ of a continuous distribution of matter in hydrodynamic motion. Thus u^μ is a unit time-like vector. Some of the following formulae are valid for any arbitrary four-vector u^μ. With the use of the covariant derivative $u_{\mu;\nu}$ one can define the following quantities which are of physical significance:

(*a*) The scalar expansion $\theta = u^\mu{}_{;\mu}$, which gives the rate at which a volume element orthogonal to the vector u^μ expands or contracts.

(*b*) A measure of the departure of the velocity field from geodesic motion is given by the acceleration $\dot{u}_\mu = u_{\mu;\nu}u^\nu$. In the absence of non-gravitational forces, such as in the case of dust (pressure-less matter), the particles follow geodesics and the acceleration vanishes.

(c) The shear tensor is symmetric, trace-free and is orthogonal to the vector u_μ. It describes the manner in which a volume element orthogonal to u^μ changes its shape, and is given as follows:

$$\sigma_{\mu\nu} = \tfrac{1}{2}(u_{\mu;\nu} + u_{\nu;\mu}) - \tfrac{1}{3}(g_{\mu;\nu} - u_\mu u_\nu)\dot{\theta} - \tfrac{1}{2}(\dot{u}_\mu u_\nu + \dot{u}_\nu u_\mu). \tag{7.12}$$

(*d*) A measure of the amount of rotational motion present in the matter is given by the vorticity tensor defined as follows:

$$w_{\mu\nu} = \tfrac{1}{2}(u_{\mu;\nu} - u_{\nu;\mu}) - \tfrac{1}{2}(\dot{u}_\mu u_\nu - \dot{u}_\nu u_\mu). \tag{7.13}$$

One can also define a vorticity vector w^μ as follows:

$$w^\mu = \tfrac{1}{2}\varepsilon^{\mu\nu\rho\sigma}u_\nu u_{\rho;\sigma}, \tag{7.14}$$

where $\varepsilon^{\mu\nu\rho\sigma}$ is the Levi–Civita alternating tensor which is antisymmetric in any pair of indices with $\varepsilon^{0123}=(-g)^{-1/2}$, g being the determinant of the metric. If the vorticity vector or tensor vanishes, the vector u^μ is said to be hypersurface orthogonal and this implies the absence of rotation in some invariant sense (rotation of the local rest frame relative to the compass of inertia; see, for example, Synge, 1937; Gödel, 1949).

Next we use (2.12) with u_μ instead of A_μ and make slight changes in the indices to get the following equation:

$$u^\mu{}_{,\alpha;\beta}-u^\mu{}_{;\beta;\alpha}=R^\mu{}_{\nu\beta\alpha}u^\nu. \tag{7.15}$$

In this equation we set μ equal to β and multiply the resulting equation with u^α as follows:

$$u^\alpha(u^\mu{}_{;\alpha;\mu}-u^\mu{}_{;\mu;\alpha})=R_{\nu\alpha}u^\nu u^\alpha, \tag{7.16}$$

where we have used (2.16). From the Einstein equation (2.22) with (2.23) we readily get

$$R_{\mu\nu}=\frac{8\pi G}{c^4}[(\varepsilon+p)u_\mu u_\nu+\tfrac{1}{2}(p-\varepsilon)g_{\mu\nu}], \tag{7.17}$$

whence it follows:

$$R_{\mu\nu}u^\mu u^\nu=\frac{4\pi G}{c^4}(\varepsilon+3p). \tag{7.18}$$

One can use the definitions of expansion, shear, vorticity and acceleration given above to write (7.16) as follows:

$$\theta_{,\alpha}u^\alpha+\tfrac{1}{3}\theta^2-\dot{u}^\alpha{}_{;\alpha}+2(\sigma^2-w^2)=-R_{\mu\nu}u^\mu u^\nu. \tag{7.19}$$

In deriving this relation the following equations have been used (the first one follows by taking the dot-derivative of $u^\mu u_\mu=1$);

$$\dot{u}_\mu u^\mu=0, \tag{7.20a}$$

$$\sigma_{\mu\nu}u^\nu=w_{\mu\nu}u^\nu=0, \tag{7.20b}$$

$$\sigma^2\equiv\tfrac{1}{2}\sigma_{\mu\nu}\sigma^{\mu\nu}, \tag{7.20c}$$

$$w^2\equiv\tfrac{1}{2}w_{\mu\nu}w^{\mu\nu}. \tag{7.20d}$$

Equation (7.19) holds for an arbitrary four-vector u^μ. We now let u^μ be the four-velocity of matter, so that (7.18) can be used in (7.19). We then get the

following important equation, known as the Raychaudhuri equation (Raychaudhuri, 1955, 1979):

$$\theta_{,\alpha}u^{\alpha}+\tfrac{1}{3}\theta^2-\dot{u}^{\alpha}_{;\alpha}+2(\sigma^2-w^2)+4\pi(\varepsilon+3p)Gc^{-4}=0. \tag{7.21}$$

The importance of this equation derives from the fact that in one form or another it is used in most if not all singularity theorems of general relativity. To see the relevance of this equation to the question of singularities we consider a simple and somewhat crude analysis. Consider a set of time-like geodesics described by the four-vector u^{μ}. Let these geodesics be irrotational. Thus we have $\dot{u}^{\mu}=w=0$. Let λ be a parameter along a typical geodesic so that $u^{\mu}=\mathrm{d}x^{\mu}/\mathrm{d}\lambda$. Then

$$\theta_{,\alpha}u^{\alpha}=\frac{\partial\theta}{\partial x^{\alpha}}\frac{\mathrm{d}x^{\alpha}}{\mathrm{d}\lambda}=\frac{\mathrm{d}\theta}{\mathrm{d}\lambda}=-\tfrac{1}{3}\theta^2-2\sigma^2-4\pi(\varepsilon+3p)Gc^{-4}. \tag{7.22}$$

Now make the assumption that $2\sigma^2+4\pi(\varepsilon+3p)Gc^{-4}$ is greater than a positive constant $\frac{1}{3}\xi^2$. Then the behaviour of θ is governed by the following differential equation:

$$\mathrm{d}\theta/\mathrm{d}\lambda=-\tfrac{1}{3}(\theta^2+\xi^2), \tag{7.23}$$

which has the solution

$$\theta=\theta_0-\xi\tan[(\xi/3)(\lambda-\lambda_0)], \tag{7.24}$$

θ_0 being the value of θ at $\lambda=\lambda_0$. From this equation it is clear that θ becomes infinite as λ is decreased from the value λ_0 to $\lambda_0-3\pi/2\xi$. If, for example, λ denotes the proper time along the geodesic, then this shows that at a finite time in the past the expansion θ becomes infinite. An infinite value of θ indicates that at that point geodesics cross each other and there is a sort of 'explosion' like the big bang. In the Friedmann models u^{μ} is given by the vector (1, 0, 0, 0) and it is readily verified that θ, which is the covariant divergence of this vector, is given by $3\dot{R}/R$. In the case $k=0$, for example, from (4.2b) we see that this is proportional to $\varepsilon^{1/2}$. We know that this tends to infinity as the big bang $t=0$ is approached. Thus the expansion θ tends to infinity at a finite time in the past. The assumption $2\sigma^2+4\pi Gc^{-4}(\varepsilon+3p)=\frac{1}{3}\xi^2$ is a limiting case. If $2\sigma^2+4\pi Gc^{-4}(\varepsilon+3p)>\frac{1}{3}\xi^2$ the infinity in θ occurs at a shorter distance away from $\lambda=\lambda_0$.

The above somewhat crude analysis can be made more precise, and this is essentially what is done in the singularity theorems. These theorems are very technical and need a great deal of preliminary apparatus. We shall here give only the statement of one of these theorems, but we need some familiarity with singularities.

7.4 Definition of singularities

The question of a definition of singularities in general relativity is a highly complex one and we can only consider a bare outline of the extensive literature on the subject. An excellent account of this topic is given in Hawking and Ellis (1973).

We have encountered a simple case of a singularity in the Friedmann models, where at $t=0$ the mass-energy density goes to infinity. The mass-energy density is a simple example of the so-called 'curvature scalars' or 'curvature invariants' whose values do not change under a coordinate transformation, so that if they are infinite at a certain point in one coordinate system, they will be infinite at that point in every coordinate system. Another example of a curvature scalar is the Ricci scalar defined by (2.20). It is well known that in empty space (where the Ricci tensor vanishes), there are four curvature invariants, one of these being $R_{\alpha\beta\gamma\delta}R^{\alpha\beta\gamma\delta}$ (see, for example, Weinberg (1972) for a discussion of this). If one of the curvature scalars goes to infinity at a point, that point is a space-time singularity, and cannot be considered as a part of the space-time manifold, whose points are defined to be such that one can introduce a coordinate system so that the metric and its derivatives to second order are well behaved. Such points may be called 'regular' points. However, all the curvature scalars remaining finite at a point does not necessarily imply the point is regular. The usual example of this that is cited is that of the two-dimensional surface of an ordinary cone in three dimensions. The curvature scalars of this surface remain finite as one approaches the apex of the cone, but the latter is not a regular point as it is not possible to introduce any coordinate system that is well behaved at that point. On the other hand, the metric behaving badly at a point does not necessarily mean that the point is singular, because the bad behaviour may be simply due to the unsuitable nature of the coordinate system. These matters are illustrated well by the Schwarzschild metric.

The Schwarzschild solution is given as follows:

$$\mathrm{d}s^2 = c^2(1-2m/r)\,\mathrm{d}t^2 - (1-2m/r)^{-1}\,\mathrm{d}r^2 - r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2). \quad (7.25)$$

Here the coefficient of $\mathrm{d}t^2$ goes to infinity at $r=0$ and that of $\mathrm{d}r^2$ goes to infinity at $r=2m$. The curvature invariants are well behaved at $r=2m$, but some of them go to infinity at $r=0$. Thus the bad behaviour of the metric cannot be removed at $r=0$, so the latter is a singularity. However, as mentioned earlier, the fact that the curvature invariants are regular at $r=2m$ does not necessarily mean that the latter is not a singularity. To prove this

one would have to find a coordinate system which is well behaved at the point. For a long time after the Schwarzschild solution was discovered, in 1916, such a coordinate system could not be found. It was observed that the radial time-like and null geodesics displayed no unusual behaviour at $r=2m$. Finally, in 1960 Kruskal found the following transformation from (r, t) to new coordinates (u, v) which shows that the point $r=2m$ is regular:

$$u^2-v^2=(2m)^{-1}(r-2m)\exp(r/2m),\ v=u\tanh(ct/4m), \tag{7.26}$$

with the metric (7.25) given as follows:

$$\mathrm{d}s^2=r^{-1}(32m^3)\exp(-r/2m)(\mathrm{d}u^2-\mathrm{d}v^2)-r^2(\mathrm{d}\theta^2+\sin^2\theta\,\mathrm{d}\phi^2), \tag{7.27}$$

where r is to be interpreted as a function of u and v given implicitly by the first equation in (7.26).

Another aspect of the question of singularities can be illustrated with the Schwarzschild metric, as follows (Raychaudhuri, 1979, p. 146). Transform the coordinate r in (7.25) to a new coordinate r' given by

$$r-2m=r'^2. \tag{7.28}$$

This changes (7.25) to the following form:

$$\begin{aligned}\mathrm{d}s=c^2r'^2/(r'^2+2m)\,\mathrm{d}t^2-(r'^2+2m)(\mathrm{d}\theta^2+\sin^2\theta\,\mathrm{d}\phi^2)\\ -4(r'^2+2m)\,\mathrm{d}r'^2.\end{aligned} \tag{7.29}$$

Clearly this metric is regular for all values of r' in $0\leqslant r'\leqslant\infty$. But this is only a part of the space represented by (7.25) with $0\leqslant r\leqslant\infty$. In (7.29) there would be no singularities of the curvature scalars such as $R_{\alpha\beta\gamma\delta}R^{\alpha\beta\gamma\delta}$ for any values of r'. It is thus not always satisfactory simply to see if the metric components are regular. One way to demand regularity which is physically meaningful is to require that all time-like and null geodesics should be complete in the sense that they can be extended to arbitrary values of their affine parameters. Since time-like and null geodesics give respectively the paths of freely falling (that is, in motion under purely gravitational forces) massive and massless particles, this requirement means that the space-time must contain complete histories of such freely falling particles, and that these geodesics should not suddenly come to an end at any point. In fact even this may not be satisfactory as the definition of a regular space-time, as Geroch (1967) has provided an example of a space-time that is geodesically complete (that is, the geodesics can be extended arbitrarily) but one that has a non-geodetic time-like curve (for example an observer propelled by a space-ship, that is, non-gravitational forces) with bounded acceleration which has a finite length. To get over these

kinds of difficulties a modified definition of completeness, called *b*-completeness, has been given by Schmidt (1973).

7.5 An example of a singularity theorem

As indicated earlier, there are various forms of singularity theorems, mostly due to Penrose, Hawking and Geroch (see Hawking and Ellis, 1973), which involve elaborate conditions, some of which are quite technical. Roughly speaking, these theorems show that quite reasonable assumptions lead to at least one consequence which is physically unacceptable. We will give here the statement of one of these theorems, due to Hawking and Penrose (1970), which is as follows:

Space-time is not time-like and null geodesically complete if:

(*a*) $R_{\mu\nu}K^\mu K^\nu \geqslant 0$ for every non-space-like vector K^μ. If the Einstein equations (2.22) are valid, and if K^μ is taken to be a unit time-like vector, this condition is readily seen to imply $T_{\mu\nu}K^\mu K^\nu \geqslant \frac{1}{2}T$. If, in addition, $T_{\mu\nu}$ is that for a perfect fluid given by (2.23) and K^μ is taken to be the four-velocity u^μ, then this condition implies $\varepsilon + 3p \geqslant 0$. For this reason this is sometimes referred to as the energy condition. Physically it is very reasonable.

(*b*) Every non-space-like geodesic contains a point at which

$$K_{[\mu}R_{\nu]\lambda\sigma[\rho}K_{\tau]}K^\lambda K^\sigma \neq 0, \qquad [\,] \text{ implies antisymmetrization,}$$

where K_μ is the tangent vector to the geodesics. This is one of the rather technical conditions and it appears that this is true for any general solution of Einstein's equations.

(*c*) There are no closed time-like curves. Physically this means that no observer can go to his past.

(*d*) There exists a point p such that the future or past null geodesics from p are focussed by the matter or curvature and start to reconverge. Penrose and Hawking show that observations on the microwave background radiation indicate that this condition is satisfied.

There are actually two alternatives to the condition (*d*) which are more technical. We refer the interested reader to Hawking and Ellis (1973, p. 266) for an account of this. We thus see that assumptions which are quite reasonable lead to consequences which are physically very strange, such as a particle's worldline suddenly coming to an end, or an observer meeting his past.

7.6 An anisotropic model

To see an example of singularities which is different from the simple Friedmann cases and yet not too complicated, we will consider in this section a model that is homogeneous but anisotropic. It is, in fact, the metric of (3.15) with $A=1$, and we use X^2, Y^2, Z^2 instead of B, C, D in that equation, so that our metric is as follows:

$$\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 - X^2(t)\,\mathrm{d}x^2 - Y^2(t)\,\mathrm{d}y^2 - Z^2(t)\,\mathrm{d}z^2. \tag{7.30}$$

This metric belongs to Bianchi type I mentioned in Section 7.2. Such models have been studied by Raychaudhuri (1958), Schücking and Heckmann (1958) and others. The case $X=Y$ with dust was considered by Thorne (1967). An account of this model is given in Hawking and Ellis (1973, p. 142).

The fact that the metric (7.30) is homogeneous has been shown at the end of Section 3.1. It is anisotropic because not all directions from a point are equivalent. There are several reasons for studying anisotropic universes. We have mentioned earlier that the universe displays a high degree of isotropy in the present epoch. However, in earlier epochs, perhaps very early ones, there may have been a significant amount of anisotropy. Also, in a realistic situation the singularity in the universe is unlikely to possess the high degree of symmetry that the Friedmann models have. The observed isotropy of the universe needs to be explained and, in the process of seeking this explanation, one must consider more general models of the universe than the Friedmann ones.

We will consider solutions of Einstein's equations for the metric (7.30) for a perfect fluid with zero pressure, that is, dust. We set $G=1$ and $c=1$ for this section and the next, and define a function $S(t)$ by $S^3=XYZ$. A solution of Einstein's equation is given as follows (M, a, b are constants):

$$\left.\begin{aligned} &\varepsilon = 3M/(4\pi S^3),\ X = S(t^{2/3}/S)^{2\sin a},\ Y = S(t^{2/3}/S)^{2\sin(a+2\pi/3)},\\ &Z = S(t^{2/3}/S)^{2\sin(a+4\pi/3)},\ S^3 = \tfrac{9}{2}Mt(t+b). \end{aligned}\right\} \tag{7.31}$$

The constant b determines the amount of anisotropy, the value $b=0$ giving the isotropic Einstein–de Sitter universe (see (4.24)). The constant 'a' determines the direction of most rapid expansion, the domain of 'a' being $-\pi/6 < a < \pi/2$. We have

$$\dot{S}/S = (2/3t)(t+\tfrac{1}{2}b)/(t+b),\ \dot{X}/X = (2/3t)[t+\tfrac{1}{2}b(1+2\sin a)]/(t+b), \tag{7.32}$$

the expressions for $\dot{Y}/Y$ and $\dot{Z}/Z$ being obtained by replacing a in $\dot{X}/X$ by $a+2\pi/3$ and $a+4\pi/3$ respectively. This universe has a highly anisotropic

singular state at $t=0$. For large t it tends to isotropy, in fact to the Einstein–de Sitter universe.

Suppose we follow the time t backwards to the initial singularity. At first there is isotropic contraction. Let $a \neq \frac{1}{2}\pi$. Then $1+2\sin(a+\frac{4}{3}\pi)$ is negative. Thus the collapse in the z-direction halts and is replaced by expansion, the rate of which becomes infinite as t tends to zero. The collapse is monotonic in the x- and y-directions. Consider now the situation forwards from $t=0$. The matter collapses from infinity in the z-direction, then halts and expands. In the x- and y-directions it expands monotonically. Thus we have here a cigar-shaped singularity. If one could observe the matter far back in time, one would see a maximum red-shift in the z-direction, then the red-shift would decrease to zero (corresponding to the halt), then one would get indefinitely large blue-shifts, the latter occurring in light given off by the matter near $t=0$.

The case $a=\frac{1}{2}\pi$ is somewhat different. Here we have

$$\dot{X}/X=(2/3t)(t+\tfrac{3}{2}b)/(t+b),\ \dot{Y}/Y=\dot{Z}/Z=(2/3)(t+b)^{-1}. \tag{7.33}$$

Following time backwards again, the initially isotropic contraction slows down to zero in the y- and z-directions but the collapse is monotonic in the x-direction. Going forwards in time, the rate of expansion of the universe in the y- and z-directions starts from a finite value but the expansion rate in the x-direction is infinite. This is thus a 'pancake' singularity. There are limiting red-shifts in the y- and z-directions, but no limit to the red-shifts in the x-direction.

7.7 The oscillatory approach to singularities

In this section we consider an interesting approach to singularities developed by Lifshitz and Khalatnikov (1963) and by Belinskii, Khalatnikov and Lifshitz (1970). We study one of the homogeneous spaces that were introduced in Section 7.2, namely, Bianchi type IX, whose structure constants are as follows (see (7.11)):

$$C^1_{23}=C^2_{31}=C^3_{12}=1. \tag{7.34}$$

Denoting (x^1, x^2, x^3) by (θ, ϕ, ψ), the three vectors $e^{(a)}_m$ (see (7.3) and (7.4)) can be taken as follows:

$$e^{(1)}_m=(\sin\psi, -\cos\psi\sin\theta, 0),\ e^{(2)}_m=(\cos\psi, \sin\psi\sin\theta, 0),$$
$$e^{(3)}_m=(0, \cos\theta, 1). \tag{7.35}$$

The metric (7.4) is given as follows, where we have taken $\eta_{ab}(t)$ to be diagonal and set $\eta_{11}=a^2$, $\eta_{22}=b^2$, and $\eta_{33}=c^2$.

$$\begin{aligned} ds^2 &= dt^2 - a^2(\sin\psi\, d\theta - \cos\psi \sin\theta\, d\phi)^2 \\ &\quad - b^2(\cos\psi\, d\theta + \sin\psi \sin\theta\, d\phi)^2 - c^2(\cos\theta\, d\phi + d\psi)^2. \end{aligned} \tag{7.36}$$

In the isotropic models studied in Chapter 3, near the singularity the spatial curvature term behaves as R^{-2} whereas the mass-energy density behaves as R^{-3} (for zero pressure) and as R^{-4} (for radiation). (See (4.2a)–(4.2c), (4.15) and (4.40).) Thus in the Friedmann models the curvature terms go to infinity slower than the terms arising from $T_{\mu\nu}$ and the derivatives with respect to time of the metric (that is, $\dot{R}$ terms). This kind of singularity is referred to as a velocity-dominated singularity (Eardley, Liang and Sachs, 1972). In the anisotropic models which are our concern in this section the behaviour near the singularity is dominated by curvature terms as observed by Belinskii and his coworkers and by Misner (1969) and is called the mixmaster singularity.

Thus if we are interested in the behaviour near the initial singularity for the anisotropic metric (7.36), it is sufficient to consider the empty space or vacuum Einstein equations where $T_{\mu\nu}=0$, for the terms arising from $T_{\mu\nu}$ are negligible in comparison to the other terms. The empty space Einstein equations can be written as follows:

$$(abc\dot{c})\dot{}/(abc) = (2a^2b^2c^2)^{-1}[(a^2-b^2)^2 - c^4], \tag{7.37a}$$

$$(ab\dot{c}c)\dot{}/(abc) = (2a^2b^2c^2)^{-1}[(b^2-c^2)^2 - a^4], \tag{7.37b}$$

$$(a\dot{b}bc)\dot{}/(abc) = (2a^2b^2c^2)^{-1}[(c^2-a^2)^2 - b^4], \tag{7.37c}$$

$$\ddot{a}/a + \ddot{b}/b + \ddot{c}/c = 0. \tag{7.37d}$$

Here a dot represents differentiation with respect to t. If the right hand sides in (7.37a)–(7.37c) were absent, we would get the following well-known Kasner (1921) solution (of Bianchi type I):

$$a = t^q,\ b = t^r,\ c = t^p, \tag{7.38}$$

where p, q, r are constants satisfying

$$p + q + r = p^2 + q^2 + r^2 = 1. \tag{7.39}$$

Suppose now that even when the terms on the right hand sides of (7.37a)–(7.37c) are present, there exist certain ranges of values of t for which the metric is given approximately by (7.38):

$$a \sim t^q,\ b \sim t^r,\ c \sim t^p. \tag{7.40}$$

Then from (7.37d) we get

$$p^2+q^2+r^2=p+q+r. \tag{7.41}$$

It is readily verified that not all the three expressions on the right hand sides of (7.37a)–(7.37c) can be positive, that is, one of these at least must be negative. From this it follows, substituting (7.40) into the left hand sides of (7.37a)–(7.37c), that at least one of the expressions $p(p+q+r-1)$, $q(p+q+r-1)$, $r(p+q+r-1)$ must be negative. The possibility that p, q, r are all positive with $p+q+r-1$ negative is inadmissible because it contradicts (7.41) (for in this case we must have $0<p<1$, $0<q<1$, $0<r<1$, so that $p^2<p$, $q^2<q$, $r^2<r$, and (7.41) becomes impossible). Thus at least one of the indices p, q, r is negative. This implies that the length along at least one direction shrinks while (since $p+q+r>0$ from (7.41)) the spatial volume, which is determined by the product $(abc)^2$ expands. In fact (7.37a)–(7.37c) do not allow two of the exponents p, q, r to be negative at the same time.

We suppose that p is negative and $q<r$. Then (7.40) implies that for small t, a and b can be neglected in comparison with c. We now define new dependent variables α, β, γ and a new independent variable τ by the following relations:

$$a=\exp(\alpha),\ b=\exp(\beta),\ c=\exp(\gamma);\ \mathrm{d}t/\mathrm{d}\tau=abc. \tag{7.42}$$

These transformations, together with the approximations introduced above, enable us to write (7.37a)–(7.37c) as follows:

$$\gamma''=-\tfrac{1}{2}\exp(4\gamma), \tag{7.43a}$$

$$\alpha''=\beta''=\tfrac{1}{2}\exp(4\gamma), \tag{7.43b}$$

where a prime denotes differentiation with respect to τ. Equation (7.43a) is in the form of the equation of motion of a particle which is moving in a potential well which is exponential. The 'velocity' γ' thus changes sign corresponding to a change from a region where c is decreasing to one where c is increasing. Belinskii *et al.* assume that the right hand sides of (7.37a)–(7.37c) are small enough at a certain epoch such that $p+q+r$ is nearly unity and one has the Kasner solution with

$$abc=wt,\ \tau=w^{-1}\log t+\text{constant}, \tag{7.44}$$

where w is a constant. Equations (6.43a) and (6.43b) can then be integrated as follows:

$$a^2=a_0^2[1+\exp(4pw\tau)]\exp(2qw\tau), \tag{7.45a}$$

$$b^2 = b_0^2[1 + \exp(4pw\tau)] \exp(2rw\tau), \tag{7.45b}$$

$$c^2 = 2|p|[\cosh(2wp\tau)]^{-1}, \tag{7.45c}$$

where we have chosen the integration constants so that as τ tends to infinity, a, b, c go to the assumed Kasner solution with a negative p. We get the following asymptotic values of a, b, c as τ tends to infinity and minus infinity respectively:

$$\text{As } \tau \to \infty,\ a \sim \exp(qw\tau),\ b \sim \exp(rw\tau),\ c \sim \exp(pw\tau), \tag{7.46a}$$

$$\text{As } \tau \to -\infty,\ a \sim \exp[w(q+2p)\tau],\ b \sim \exp[w(r+2p)\tau],\ c \sim \exp(-pw\tau), \tag{7.46b}$$

In (7.46a) we have $w\tau \sim \log t$ while in (7.46b), $w(1+2p)\tau \sim t$. In the second of these limits, that is in (7.46b), transforming back to t from τ (with $w(1+2p)\tau = t$), we get

$$a \sim t^{q'},\ b \sim t^{r'},\ c \sim t^{p'}, \tag{7.47}$$

where

$$p' = -p/(1+2p) > 0, \tag{7.48a}$$

$$q' = (2p+q)/(1+2p) < 0, \tag{7.48b}$$

$$r' = (r+2p)/(1+2p) > 0. \tag{7.48c}$$

This behaviour is different from that existing in the limit $\tau \to \infty$ which is given by (7.40), in the sense that the exponent in c has changed from negative to positive, while that of a has become negative (that is, q is positive but q' negative). Thus the a- and c-axes have interchanged their expanding and contracting behaviours. This indicates that, as we move towards the singularity, distances along two of the axes oscillate while that along the third axis decreases monotonically. This happens in successive periods which are called 'eras'. On going from one era to the next, the axis along which distances decrease monotonically changes to another one. Asymptotically the order in which this change occurs becomes a random process (Landau and Lifshitz, 1975). One has a particularly long era if (p, q, r) corresponds to the triplet (1, 0, 0). In this case there are no particle horizons (see Section 4.7) in the direction for which the index is unity, since $\int_0 t^{-1}\, dt$ diverges. In the course of evolution this particular direction also changes and this phenomenon may lead to effective abolition of all particle horizons. This was one of the motivations of the mixmaster model of Misner which was thought to provide the solution to the 'horizon'

problem mentioned in Chapter 1, that is, to explain why the universe is so isotropic and homogeneous. But this model did not provide a solution to the problem, although some interesting insights were gained. This completes our brief exposition of singularities in cosmology. For more details of the material presented in this chapter, we refer to the books by Hawking and Ellis (1973), Raychaudhuri (1979), Landau and Lifshitz (1975) and the papers cited in this chapter. There have also been interesting inhomogeneous exact cosmological solutions following the work of Szekeres (1975); see, for example, the papers by Szafron (1977), Szafron and Wainwright (1977), Wainwright (1979), Wainwright and Marshman (1979), Wainwright, Ince and Marshman (1979), Wainwright and Goode (1980), Wainwright (1981), and Goode and Wainwright (1982). It is however, beyond the scope of this book to consider these models.

7.8 A singularity-free universe?

A new class of inhomogeneous cosmological solutions has been found by Senovilla (1990) which does not seem to possess any singularities in the past, with the curvature and matter invariants regular and smooth everywhere. The source is a perfect fluid with equation of state $\varepsilon = 3p$. The metric is as follows (with signature $+2$):

$$\mathrm{d}s^2 = e^{2f}(-\mathrm{d}t^2 + \mathrm{d}x^2) + K(q\,\mathrm{d}y^2 + q^{-1}\,\mathrm{d}z^2), \tag{7.49}$$

where the functions f, K and q depend on t and x only and are given explicitly as follows:

$$\begin{aligned} e^f &= [A\cosh(at) + B\sinh(at)]^2\cosh(3ax), \\ K &= [A\cosh(at) + B\sinh(at)]^2\sinh(3ax)[\cosh(3ax)]^{-3/2}, \\ q &= [A\cosh(at) + B\sinh(at)]^2\sinh(3ax), \end{aligned} \tag{7.50}$$

where a, A, B are arbitrary constants. The pressure and energy density are given as follows:

$$p = \tfrac{1}{3}\varepsilon = 5\chi^{-1}a^2[A\cosh(at) + B\sinh(at)]^{-4}[\cosh(3ax)]^{-4}, \tag{7.51}$$

where χ is the gravitational constant in suitable units.

In two important papers, Raychaudhuri (1998, 1999) evaluates the new Senovilla solution and re-examines the singularlity theorems, and offers an additional theorem. To recapitulate, there are essentially four conditions: (1) the causality condition forbidding closed time-like lines, (2) the strong energy condition $(T_{\mu\nu} - \frac{1}{2}g_{\mu\nu}T)u^\mu u^\nu \geqslant 0$, (3) a condition on the Riemann–Christoffel tensor, and (4) existence of a trapped surface. Raychaudhuri

quotes from Misner, Thorne and Wheeler (1973): 'All the conditions except the trapped surface seem eminently reasonable for any physically realistic space time' (p. 935). Raychaudhuri also discusses the further solutions found by Ruiz and Senovilla (1992). One of the important points to notice is that it is the last condition that is violated by the new singularity-free solution. However, as Raychaudhuri shows, the average of the physical and kinematic scalars taken over the entire space-time vanishes. In the new solution the space-time is open in all directions, which means, according to Raychaudhuri, that the space-time has topology $R^3 \times R$. Raychaudhuri goes on to enunciate and prove an interesting new theorem: 'In any singularity free non-rotating universe, open in all directions, the space-time average of all stress energy invariants including the energy density vanishes.' Here 'non-rotating' means all matter has worldlines forming a normal congruence, that is, one that is hypersurface orthogonal. This means essentially that the tangent four-vectors to the worldlines are orthogonal to the space-like three-surface on which the matter lies at any instant. The proof is based on Raychaudhuri's earlier equation (7.21). He goes on to discuss interesting implications.

8

The early universe

8.1 Introduction

As mentioned in Chapter 1, the 'cosmic background radiation' discovered originally by Penzias and Wilson in 1965 provides evidence that the universe must have gone through a hot dense phase. We have also seen that the Friedmann models (described in Chapter 4), if they are regarded as physically valid, predict that the density of mass-energy must have been very high in the early epochs of the universe. In fact, of course, the Friedmann models imply that the mass-energy density goes to infinity as the time t approaches the 'initial moment' or 'the initial singularity', at $t=0$. This is what is referred to as the 'big bang', meaning an explosion at every point of the universe in which matter was thrown asunder violently, from an infinite or near infinite density. However, the precise nature of the physical situation at $t=0$, or the situation *before* $t=0$ (or whether it is physically meaningful to talk about any time *before* $t=0$) – these sorts of questions are entirely unclear. In this and the following chapter we shall try to deal partly with some questions of this kind. In the present chapter we simply *assume* that there was a catastrophic event at $t=0$, and try to describe the state of the universe from about $t=0.01$ s until about $t=$ one million years. This will be our definition of the 'early universe', which specifically excludes the first hundredth of a second or so, during which, as we shall see in the next chapter, and as speculations go, events occurred which are of a very different nature from those occurring in the 'early universe' according to the definition given here.

In this section we shall describe qualitatively the state of the early universe and in the following sections we shall provide a more quantitative account of this state. The description given in this section is derived largely

from that given in Weinberg's book (1977, 1983). As indicated in Fig. 1.3, the spectrum of the cosmic background radiation peaks at slightly under 0.1 cm. Penzias and Wilson made their original observation at 7.35 cm. Since that time there have been many observations, both ground-based and above the atmosphere, which confirm the black-body nature of the radiation, with a temperature of about 2.7 K. Below about 0.3 cm, the atmosphere becomes increasingly opaque, so such observations have to be carried out above the atmosphere. Although at times there have been slight doubts, it is now generally agreed that the cosmic background radiation is indeed the remnant of the radiation from the early universe, which has been red-shifted, that is, reduced in temperature to 2.7 K. As we shall see later in more detail, the temperature of the cosmic background radiation provides us with an important datum about the universe, that there are about 1000 million photons in the universe for every nuclear particle; by the latter we mean protons and neutrons, or 'baryons'. There is some uncertainty in this figure, but we shall use this figure for the time being, and later explain the possible modification.

To describe the state of the early universe we choose several instants of time, which are referred to by Weinberg as 'frames', as if a movie had been made and we were looking at particular frames in this movie. These instants of time are chosen so that major changes take place near those times. In the following we describe the physical state of the universe at these instants, or frames. (The values of the temperature, time, etc., are slightly different from those in Weinberg (1977, 1983) to conform with subsequent calculations in this book.)

(i) First frame

This is at $t = 0.01$ s, when the temperature is around 10^{11} K, which is well above the threshold for electron–positron pair production. The main constituents of the universe are photons, neutrinos and antineutrinos, and electron–positron pairs. There is also a small 'contamination' of neutrons, protons and electrons. The energy density of the electron–positron pairs is roughly equal to that of the neutrinos and antineutrinos, both being $\frac{7}{4}$ times the energy density of the photons. The total energy density is about 21×10^{44} eV 1^{-1}, or about 3.8×10^{11} g cm^{-3}. The characteristic expansion time of the universe (that is, the reciprocal of Hubble's 'constant' at that instant, which is the age of the universe if the rate of expansion had been the same from the beginning as at that instant) is 0.02s. The neutrons and protons cannot form into nuclei, as the latter are unstable. The spatial volume of the universe would be either infinite or, if it is one of the finite

models, say with density twice the critical density, its circumference would be about 4 light years.

(ii) Second frame

This is at $t = 0.12$ s, when the temperature has dropped to about 3×10^{10} K. No qualitative changes have occurred since the first frame. As in the first frame, the temperature is above electron–positron pair threshold, so that these particles are relativistic, and the whole mixture behaves more like radiation than matter, with the equation of state given nearly by $p = \frac{1}{3}\varepsilon$. The total density is about 3×10^7 g cm^{-3}. The characteristic expansion time is about 0.2 s. No nuclei can be formed yet, but the previous balance between the numbers of neutrons and protons, which were being transformed into each other through the reaction $n + \nu \rightleftarrows p + e^-$, is beginning to be disturbed as neutrons now turn more easily into the lighter protons than vice versa. Thus the neutron–proton ratio becomes approximately 38% neutrons and 62% protons. The thermal contact (see below) between neutrinos and other forms of matter is beginning to cease.

(iii) Third frame

This is at $t = 1.1$ s, when the temperature has fallen to about 10^{10} K. The thermal contact between the neutrinos and other particles of matter and radiation ceases. Thermal contact is here taken to mean the conversion of electron–positron pairs into neutrino–antineutrino pairs and vice versa, the conversion of neutrino–antineutrino pairs into photons and vice versa, etc. Henceforth neutrinos and antineutrinos will not play an active role, but only provide a contribution to the overall mass-energy density. The density is of the order of 10^5 g cm^{-3} and the characteristic expansion time is a few seconds. The temperature is near the threshold temperature for electron–positron pair production, so that these pairs are beginning to annihilate more often to produce photons than their creation from photons. It is still too hot for nuclei to be formed and the neutron–proton ratio has changed to approximately 24% neutrons and 76% protons.

(iv) Fourth frame

This is approximately at $t \approx 13$ s, when the temperature has fallen to about 3×10^9 K. This temperature is below the threshold for electron–positron production so most of these pairs have annihilated. The heat produced in this annihilation has temporarily slowed down the rate of cooling of the universe. The neutrinos are about 8% cooler than the photons, so the energy density is a little less than if it were falling simply as the fourth

power of the temperature (recall that according to the Stefan–Boltzmann law $\varepsilon = \sigma T^4$ erg cm^{-3}, where $\sigma = 7.564\ 64 \times 10^{-15}$, and T is the temperature in K). The neutron–proton balance has shifted to about 17% neutrons and 83% protons. The temperature is low enough for helium nuclei to exist, but the lighter nuclei are unstable, so the former cannot be formed yet. By helium nuclei we mean alpha particles, He^4, which have two protons and two neutrons. The expansion rate is still very high, so only the light nuclei form in two-particle reactions, as follows: $p + n \rightarrow D + \gamma$, $D + p \rightarrow He^3 + \gamma$, $D + n \rightarrow H^3 + \gamma$, $He^3 + n \rightarrow He^4 + \gamma$, $H^3 + p \rightarrow He^4 + \gamma$. Here D denotes deuterium, which has one neutron and one proton, He^3 is helium-3, an isotope of helium with two protons and one neutron, H^3 is tritium, an isotope of hydrogen with one proton and two neutrons, and γ stands for one or more photons. Although helium is stable, the lighter nuclei mentioned here are unstable at this temperature, so helium formation is not yet possible, as it is necessary to go through the above intermediate steps to form helium. The energy required to pull apart the neutron and proton in a D nucleus, for example, is one-ninth that required to pull apart a nucleon (neutron or proton) from an He^4 nucleus. In other words, the binding energy of a nucleon in deuterium is one-ninth that in an He^4 nucleus.

(v) Fifth frame

This is about 3 min after the first frame when the temperature is about 10^9 K, which is approximately 70 times as hot as the centre of the Sun. The electron–positron pairs have disappeared, and the contents of the universe are mainly photons and neutrinos plus, as before, a 'contamination' of neutrons, protons and electrons (whose numbers are much smaller than the number of photons, by a ratio of about $1{:}10^9$), which will eventually turn into the matter of the present universe. The temperature of the photons is about 35% higher than that of the neutrinos. It is cool enough for H^3, He^3 and He^4 nuclei to be stable, but the deuterium 'bottleneck' is still at work so these nuclei cannot be formed yet. The beta decay of the neutron into a proton, electron and antineutrino is becoming important, for this reaction has a time scale of about 12 min. This causes the neutron–proton balance to become 14% neutrons and 86% protons.

A little later than the fifth frame the temperature drops enough for deuterium to become stable, so that heavier nuclei are quickly formed, but as soon as He^4 nuclei are formed other bottlenecks operate, as there are no stable nuclei at that temperature with five or eight particles. The exact temperature depends on the number of photons per baryon; if this number is 10^9 as assumed before, then the temperature is about 0.9×10^9 K, and these

events take place at some time between $t = 3$ min and $t = 4$ min. Nearly all the neutrons are used up to make He^4, with very few heavier nuclei due to the other bottlenecks mentioned. The neutron–proton ratio is about 12% or 13% neutrons to 88% or 87% protons, and it is frozen at this value as the neutrons have been used up. As the He^4 nuclei have equal numbers of neutrons and protons, the proportion of helium to hydrogen nuclei (the latter being protons) by weight is about 24% or 26% helium and 76% or 74% hydrogen. This process, by which heavier nuclei are formed from hydrogen, is called *nucleosynthesis*. If the number of photons per baryon is lower (that is, if the baryon: photon ratio is higher), then nucleosynthesis begins a little earlier, and slightly more He^4 nuclei are formed than 24% or 26% by weight.

(vi) Sixth frame

This is approximately at $t \simeq 35$ min, when the temperature is about 3×10^8 K. The electrons and positrons have annihilated completely, except for the small number of electrons left over to neutralize the protons. It is assumed throughout that the charge density in any significant volume of the universe is zero. The temperature of the photons is about 40% higher than the neutrino temperature, and will remain so in the subsequent history of the universe. The energy density is about 10% the density of water, of which 31% or so is contributed by neutrinos and the rest by photons. The density of 'matter' (that is, of the nuclei and protons, etc.) is negligible in comparison to that of photons and neutrinos. The characteristic expansion time of the universe is about an hour and a quarter. Nuclear processes have then stopped, the proportion of He^4 nuclei being anywhere between 20% and 30% depending on the baryon : photon ratio (see Fig. 8.1).

We see from the preceding discussion that the proportion of helium nuclei formed in the early universe was anything from 20–30% by weight, with very few heavier nuclei due to the five- and eight-particle bottlenecks. For the nucleosynthesis process to take place one needs temperatures of the order of a million degrees. After the temperature dropped below about a million degrees in the early universe, the only place in the later universe where similar temperatures exist would be the centre of stars. It can be shown that no significant amount of helium (compared to the 20–30% of the early universe) could have been created in the centre of stars. This follows from the fact that such a significant amount of helium formation would have released so much energy into the interstellar and intergalactic space, that it would be inconsistent with the amount of radiation actually

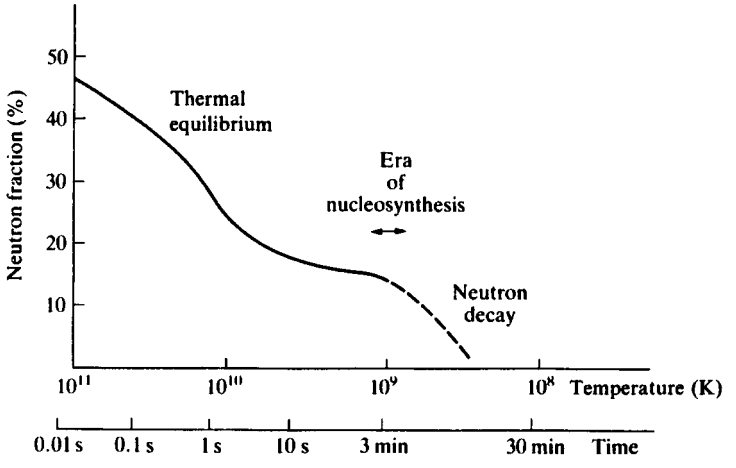


Fig. 8.1. Diagram to describe the neutron–proton ratio in the early universe. The period of 'thermal equilibrium' is one in which all particles and radiation are in equilibrium and the neutron–proton ratio depends on the mass difference between these particles. The 'era of nucleosynthesis' is the period when lighter nuclei, predominantly helium, are being formed. The dashed portion indicates that if the neutrons had not been incorporated into nuclei they would have decayed through beta decay (Weinberg, 1977).

given off since the time of star and galaxy formation, an amount of which can be calculated from the average absolute luminosity of stars and galaxies, which are known, and the time scale during which these have existed, which is from soon after the recombination era (see below). Thus if the above picture is reasonable, there should be approximately 20–30% helium nuclei in the present universe, most of the rest being predominantly hydrogen, with a small amount of heavier nuclei. This is indeed found to be the case. We shall have more to say about this later in this chapter.

We have seen that the time, temperature and the extent of nucleosynthesis depends on the density of nuclear particles compared to photons. The amount of deuterium that was produced by nucleosynthesis in the early universe, and the amount that survives and should be observable today, depends very sensitively on the nuclear particle to photon ratio. As an illustration of this, we give in Table 8.1 the abundance of deuterium as worked out by Wagoner (1973) for three values of the photon : nuclear particle ratio. We shall have more to say about deuterium later in this chapter.

We have seen that after the first few minutes the only particles left in the

Table 8.1 *Abundance of deuterium and the photon : baryon ratio.*

Photons: nuclear particle	Deuterium abundance (parts/10^6)
100 million	0.000 08
1 000 million	16
10 000 million	600

universe were photons, neutrinos, neutrons, protons and electrons. The latter two particles are charged ones, and in their free state they could scatter photons freely. As a result the 'mean free path' of photons, that is, the average distance that a photon travels in between scatterings by two charged particles, was small compared to the distance a photon would travel during the characteristic expansion time of the universe for that period, if it were unimpeded. This is what is meant by the matter and radiation being in equilibrium, as there is free exchange of energy between the two. Thus the universe, during the period that protons and electrons were free particles, was opaque to electromagnetic radiation.

Eventually the temperature of the universe was cool enough for electrons and protons to form stable hydrogen atoms in their ground state when they combined. Now it takes about 13.6 eV to ionize a hydrogen atom completely, that is, pull apart the electron from the proton. The energy of a particle in random motion at a temperature of T K is kT, where k is Boltzmann's constant. Thus the temperature corresponding to an energy of 13.6 eV is k^{-1} times 13.6, where k^{-1} is approximately 11 605 K eV^{-1}. This gives about 1.576×10^5 K as the temperature at which a hydrogen atom is completely ionized. However, even in the excited states, in which it is not ionized, a hydrogen atom can effectively scatter photons. Thus it is only in the ground state that it ceases to interact significantly with photons. The temperature at which the primeval protons and electrons combined to form the ground state hydrogen atoms was about 3000–4000 K, which occurred a few hundred thousand years after the big bang. This era is referred to as 'recombination' (a singularly inappropriate term, as Weinberg remarks, as the electrons and protons were never in a combined state before!). After this period the universe became transparent to electromagnetic radiation, that is, the mean free path of a photon became much longer than the distance traversed in a characteristic expansion time of the period. This is the reason we get light, which has hardly been impeded, except for the red-shift, from galaxies billions of light years away.

8.2 The very early universe

In the last section we discussed qualitatively the early universe which we defined to begin at about $t = 0.01$ s. In this section we shall give a qualitative and speculative discussion of the very early universe, which we take to be the first hundredth of a second or so. As mentioned in Chapter 1 and also earlier in this chapter, there have been elaborate speculations about the very early universe. We shall discuss these speculations in some detail in the next chapter, where we shall give a quantitative discussion wherever possible. Some of the remarks made in this section may have to be qualified in the next chapter.

As we shall see more clearly in the next chapter, the very early universe involves elementary particles and their interactions in an intimate way, much more so than the early universe. For this reason it is necessary to know something about these particles. Table 8.2 gives the classification and properties of the more common elementary particles. As is well known from quantum field theory, which is the theory describing the interactions of these particles, the interactions can be described in a picturesque way by Feynman diagrams, which give the amplitudes for various processes to certain order in the coupling constant. Three such diagrams are given in Fig. 8.2, corresponding to electromagnetic, strong and weak interactions. These interactions and gravitation are described in Table 8.3. There is a considerable amount of uncertainty in our knowledge of the first hundredth of a second of the universe. This stems partly from our inadequate knowledge of the strong interactions of elementary particles. As we go to higher temperatures than the first frame temperature of about 10^{11} K, nearer $t = 0$, there would be copious production of hadrons, and it becomes difficult to describe the nature of matter at these temperatures for this reason, as the hadrons take part in strong interactions, whose precise nature is not known. There are two views of the nature of matter at such energies. The first one, which is not in favour at present, says that there are no 'elementary' hadrons but that every hadron is in a sense a composite of all other hadrons. In this case, as the temperature increases, the energy available goes into producing more massive hadrons, and not into the random motion of the constituent particles. As there is no limit to the mass of these 'elementary' hadrons, there is a maximum possible temperature, around 2×10^{12} K, even though the density goes to infinity. The idea of this 'nuclear democracy' was mainly due to G. F. Chew; the maximum temperature in hadron physics was pointed out by R. Hagedorn (see, for example, Huang and Weinberg (1970)).

Table 8.2 *In this table are listed the more common elementary particles. Particles and their antiparticles have the same mass, same life-time and opposite charges, so they are listed in the same line. A symbol with a bar over it denotes an antiparticle; thus $\bar{\nu}_\mu$ is the muon-antineutrino. Leptons take part in weak interactions but not in strong interactions. All hadrons take part in strong interactions: they are made up of mesons (which are bosons) and baryons (which are fermions). All hadrons also take part in weak interactions. Baryons other than the proton and the neutron are called hyperons.*

	Particle	Symbol	Charge (in units of proton charge)	Mass (MeV)	Life-time (s)	Spin (in units of $\hbar$)
	Photon	γ	0	0	infinite	1
Leptons	Neutrino	ν_e, $\bar{\nu}_e$	0	less than 0.001	infinite (?)	$\frac{1}{2}$
		ν_μ, $\bar{\nu}_\mu$	0	less than 0.001	infinite (?)	$\frac{1}{2}$
	Electron	$e^\pm$	± 1	0.51	infinite	$\frac{1}{2}$
	Muon	$\mu^\pm$	± 1	105.66	2.2×10^{-6}	$\frac{1}{2}$
Mesons	Pion	$\pi^\pm$	± 1	139.57	2.6×10^{-8}	0
		π^0	0	134.97	0.84×10^{-16}	0
	Kaon	$\kappa^\pm$	± 1	493.71	1.24×10^{-8}	0
		κ^0, $\bar{\kappa}^0$	0	497.71	0.88×10^{-10}	0
	Eta	η	0	548.8	2.50×10^{-17}	0
Hadrons	Proton	p, $\bar{p}$	± 1	938.26	infinite (?)	$\frac{1}{2}$
	Neutron	n, $\bar{n}$	0	939.55	918	$\frac{1}{2}$
	Lambda hyperon	Λ, $\bar{\Lambda}$	0	1115.59	2.52×10^{-10}	$\frac{1}{2}$

Baryons	Sigma hyperon	$\Sigma^+, \bar{\Sigma}^+$	± 1	1189.42	8.00×10^{-11}	$\frac{1}{2}$
	Sigma hyperon	$\Sigma^0, \bar{\Sigma}^0$	0	1192.48	less than 10^{-14}	$\frac{1}{2}$
	Sigma hyperon	$\Sigma^-, \bar{\Sigma}^-$	± 1	1197.34	1.48×10^{-10}	$\frac{1}{2}$
	Cascade hyperon	$\Xi^0, \bar{\Xi}^0$	0	1314.7	2.98×10^{-10}	$\frac{1}{2}$
	Cascade hyperon	$\Xi^-, \bar{\Xi}^-$	± 1	1321.3	1.67×10^{-10}	$\frac{1}{2}$
	Omega hyperon	$\Omega^-, \bar{\Omega}^-$	± 1	1672	1.3×10^{-10}	$\frac{1}{2}$

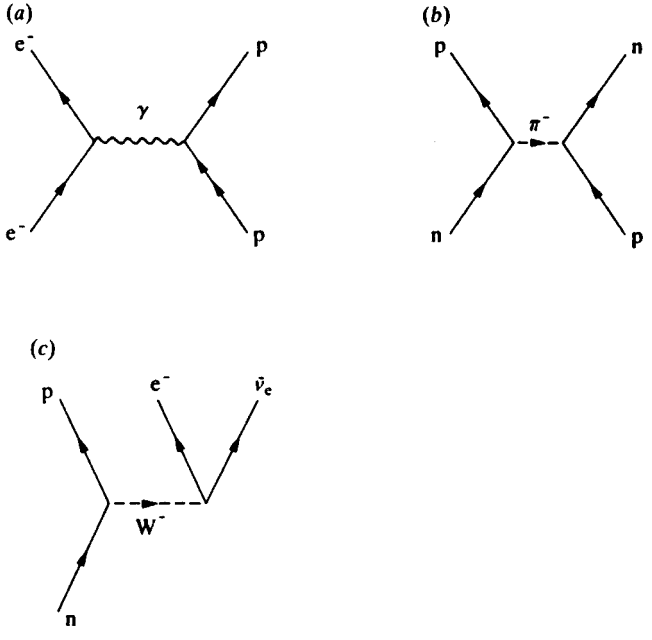


Fig. 8.2. This figure illustrates how forces are mediated by the exchange of particles. In (*a*) an electron (e^-) and a proton (p) interact by exchanging a photon (γ). In (*b*) a neutron becomes a proton by emitting a π-meson, which is then absorbed by another proton which subsequently becomes a neutron. In (*c*) the beta decay of a neutron is caused by the emission of an intermediate vector meson W^- which decays into an electron and an electron-antineutrino.

In the second view of particle physics all hadrons are made of a few fundamental constituents, known as quarks. They come in six varieties, known as 'flavours', these being the up, down, strange, charmed, top and bottom quarks, represented respectively by the letters u, d, s, c, t, b (the latter two are sometimes called 'truth' and 'beauty'). There are also the corresponding aniquarks denoted by $\bar{u}$, $\bar{d}$, etc. These quarks have fractional charges (see Table 8.4) and each flavour comes in three states called 'colours', usually referred to as yellow, blue and red, with the corresponding antiquarks being antiyellow, etc. Within a baryon or a meson the quarks interact with each other by exchanging still other fundamental particles called 'gluons' of which there are eight kinds, depending on their colour composition. The hadrons are 'colourless', being composite of quarks of all three colours, or quarks of a certain colour and its anticolour.

The Glashow–Weinberg–Salam theory gives a unified description of the

Table 8.3 *This table gives some properties of the four kinds of forces encountered in nature so far, namely gravitational, electromagnetic, strong (nuclear) and weak forces. 'Particles exchanged' means the particles through the exchange of which the corresponding force is mediated. The 'graviton' is a hypothetical particle through the exchagne of which gravitational forces are mediated.*

	Gravitational force	Electromagnetic force	Strong (nuclear) force	Weak force
Range	Infinite	Infinite	10^{-13}–10^{-14} cm	Less than 10^{-14} cm
Examples	Astronomical forces	Atomic forces	Nuclear forces	Beta decay of neutron
Strength	10^{-39}	$\frac{1}{137}$	1	10^{-5}
Particles acted upon	Everything	Charged particles	Hadrons	Hadrons and leptons
Particles exchanged	Gravitons (?)	Photons	Hadrons	Intermediate vector bosons

Table 8.4 *In one form of the grand unified theories there is a correspondence between leptons and quarks, as shown in this table. See the text for the meaning of the quark symbols. The τ^- refers to the τ-lepton and ν_τ is the corresponding neutrino. Each of the quarks come in three 'colours'.*

	Leptons		Quarks	
	Symbol	Charge	Symbol	Charge
First generation	ν_e	0	u	$+\frac{2}{3}$
	e^-	-1	d	$-\frac{1}{3}$
Second generation	ν_μ	0	c	$+\frac{2}{3}$
	μ	-1	s	$-\frac{1}{3}$
Third generation	ν_τ	0	t	$+\frac{2}{3}$
	τ	-1	b	$-\frac{1}{3}$

weak and electromagnetic interactions, according to which above a certain energy both interactions are similar and have the same strength. There have been attempts at unifying these with the strong interactions – the Grand Unified Theories – but these have not been so successful.

Although there are strong indications that hadrons are made of quarks, no quarks have been observed yet. A satisfactory explanation of this phenomenon has not been found, although there are some hints in the property of 'asymptotic freedom', which is a consequence of the gauge theory which is thought to describe the interactions of quarks and gluons. This property indicates that the strength of the interaction between two quarks becomes negligible when they are close together, and correspondingly the strength increases when they are far apart. Thus if one attempts to detach a quark from other quarks in a baryon, say, the energy required eventually becomes so great that a quark–antiquark pair is formed, so that these combine with the existing quarks to form two hadrons, and one does not get a free quark. Thus in the quark model, in the very early universe the quarks must have been very close to each other and so behaved essentially as free particles. As the universe cooled, every quark must have either annihilated with another quark to produce a meson, or else formed a part of a neutron or a proton. In this case both the temperature and the density tends to infinity as t tends to zero.

There is a possibility that the universe may have suffered a phase transition as the universe cooled, somewhat like the freezing of water. At this

phase transition, the electromagnetic and weak interactions may have become different. In the Glashow–Weinberg–Salam unification of electromagnetism and the weak interactions, the basic theory used is a gauge theory. One way of looking at this unification is as follows. Electromagnetic interactions between charged particles are mediated via the photon, which is a massless spin 1 particle (see Fig. 8.2(*a*)). The weak interactions are mediated by massive intermediate vector bosons, the $W^{\pm}$ and Z^0 particles, which are spin 1 particles with masses of about 80 and 90 proton masses respectively. Now at energies which are much higher than the energies represented by these masses, the intermediate boson masses can be neglected so that the weak interactions can be considered as being mediated by massless spin 1 particles. This is akin to the electromagnetic interactions so that at these energies the two interactions behave in a similar manner. It was shown in 1972 by Kirzhnits and Linde that, in fact, gauge theories exhibit a phase transition at a critical temperature of about 3×10^{15} K. Above this temperature the unity between the electromagnetic and weak interactions that is incorporated in the Glashow–Weinberg–Salam model was manifest. Below this temperature the weak interactions became short range while the electromagnetic interactions continued to be long range (these are characteristics of interactions which are mediated respectively by massive and massless particles). When water freezes, a certain symmetry is lost, for example, ice crystals at any point do not possess the same rotational symmetry as liquid water. Secondly, the frozen ice is separated into different domains with different crystal structures. It is conceivable that after the phase transition at some critical temperature the universe has different domains in which the erstwhile symmetry between the electromagnetic and weak interactions is broken in different manners, and that we live in one of these domains. There may remain in the universe zero-, one- or two-dimensional 'defects' from the time of the phase transition.

There is also the possibility that at higher temperatures there may have been symmetry between all three of the microscopic interactions – the weak, electromagnetic and strong interactions, and at yet higher temperatures the weakest of the forces, gravitation, may also have been included in this symmetry. At superhigh temperatures the energies of particles in thermal equilibrium may be so large that the gravitational force between them may be comparable to any other force. This may occur at 10^{32} K, at about 10^{-43} s after $t=0$. In this situation the horizon would be at a distance less than what we regard as the radius of the particles, that is, crudely speaking, each particle would be as big as the observable universe!

Just as neutrinos and then photons decoupled from matter and continued to form a part of the 'background' radiation, so at a much earlier time gravitational radiation would have also decoupled and there must also be present cosmic background gravitational radiation with a temperature of about 1 K. If it could be detected, it would give us information about a much earlier epoch of the universe than the photon or the neutrino background radiation. However, this is far beyond present technology, as gravitational radiation has not yet been detected in any form.

After the above qualitative descriptions of the early and the very early universe, we go on to more quantitative descriptions in this and the next chapter. The rest of this chapter is based mainly on Weinberg (1972, 1983), Schramm and Wagoner (1974), Bose (1980), and Gautier and Owen (1983).

8.3 Equations in the early universe

We see from (4.15) and (4.40) that in the matter-dominated and radiation-dominated situations the mass-energy density varies as R^{-3} and R^{-4} respectively. Thus in these situations εR^2 varies as R^{-1} and R^{-2} respectively. We know that in all the Friedmann models R starts from the value zero at $t=0$. Thus in any case εR^2 tends to infinity as t tends to zero. This shows (see (3.76a) and (4.2a)–(4.2c)) that near $t=0$, that is, in the early universe, one can approximate the evolution of R for all three values of k by the same equation, (4.2b), that is, as follows:

$$\dot{R}^2=(8\pi G/3)\varepsilon R^2/c^2. \tag{8.1}$$

This in turn shows that the initial behaviour of R is independent of whether the universe is open or closed. We have seen that the early universe is dominated either by radiation or radiation and highly relativistic particles. For these the equation of state is $p=\frac{1}{3}\varepsilon$, so that we get the mass-energy density ε behaving as R^{-4}. Now according to the Stefan–Boltzmann law the energy density of radiation varies as T^4, where T is the absolute temperature. Thus the temperature of the radiation (and relativistic matter) in the early universe varies as R^{-1}. After the decoupling of matter and radiation the temperature of the radiation continues to decrease as R^{-1}. For a short period there is modification of this behaviour (see below).

Equation (8.1) has the consequence that in the early universe R behaves as $t^{1/2}$, since the equation of state is $p=\frac{1}{3}\varepsilon$ (see (4.47)). If the early universe had been matter-dominated, R would have varied as $t^{2/3}$ (see (4.45)).

The early universe, which is radiation-dominated, can thus be characterized by connecting values of R, ε, T at any two instants of time t_1 and t_2, as follows:

$$R_1/R_2 = t_1^{1/2}/t_2^{1/2} = \varepsilon_2^{1/4}/\varepsilon_1^{1/4} = T_2/T_1, \tag{8.2}$$

provided no major changes take place in the constitution of the contents, such as electron–positron annihilation. For example, for the whole of the radiation-dominated period after the electron–positron annihilation, the energy density is given as follows:

$$\varepsilon = 1.22 \times 10^{-35} T^4 \text{ g cm}^{-3}, \tag{8.3}$$

(see (8.23) below) where here, as elsewhere, T denotes absolute temperature.

8.4 Black-body radiation and the temperature of the early universe

Although the properties of black-body radiation are well known, we give here a brief summary for completeness. The energy density of black-body radiation in a range of wavelengths from λ to $\lambda + \mathrm{d}\lambda$ is given by the Planck formula as follows:

$$\mathrm{d}u = (8\pi hc/\lambda^5)\, \mathrm{d}\lambda [\exp(hc/kT\lambda) - 1]^{-1}, \tag{8.4}$$

where k is Boltzmann's constant (1.38×10^{-16} erg K^{-1}), h is Planck's constant (6.625×10^{-27} erg s). For long wavelengths, neglecting higher powers of λ^{-1}, (8.4) reduces to

$$\mathrm{d}u = (8\pi kT/\lambda^4)\, \mathrm{d}\lambda, \tag{8.5}$$

which is the Rayleigh–Jeans formula. If this formula is continued to $\lambda = 0$, one gets an infinite energy density. The maximum of $\mathrm{d}u$ in the Planck formula (8.4) occurs at the value of λ given by the following equation:

$$5kT\lambda[\exp(hc/kT\lambda) - 1] = hc \exp(hc/kT\lambda). \tag{8.6}$$

The solution of this transcendental equation is given approximately as follows:

$$\lambda_0 = 0.201\,405\,2hc/kT, \tag{8.7}$$

which shows that the wavelength at which the maximum occurs is inversely proportional to the temperature. The total energy at temperature T is obtained by integrating (8.4) over all wavelengths:

$$u = \int_0^\infty (8\pi hc/\lambda^5)[\exp(hc/kT\lambda) - 1]^{-1}\, \mathrm{d}\lambda$$
$$= \int_0^\infty (8\pi h\nu^3/c^3)[\exp(h\nu/kT) - 1]^{-1}\, \mathrm{d}\nu, \tag{8.8}$$

where in the last step we have expressed the integral in terms of the frequency $\nu = c/\lambda$. The result of the integration is as follows:

$$u = 8\pi^5(kT)^4/15h^3c^3 = 7.5641 \times 10^{-15}\ T^4 \text{ erg cm}^{-3}. \tag{8.9}$$

Since a photon has energy $h\nu = hc/\lambda$, the number density of photons is given as follows, for wavelengths from λ to $\lambda + \mathrm{d}\lambda$:

$$\mathrm{d}N = \mathrm{d}u/h\nu = \lambda\, \mathrm{d}u/hc = (8\pi/\lambda^4)[\exp(hc/kT\lambda) - 1]^{-1}, \tag{8.10}$$

and the number density of photons is

$$N = \int_0^\infty \mathrm{d}N = 60.421\ 98(kT)^3/(hc)^3 = 20.28T^3 \text{ photons cm}^{-3} \tag{8.11}$$

and the energy per photon is

$$u/N = 3.73 \times 10^{-16}T. \tag{8.12}$$

Equation (8.11) enables us to make a rough estimate of the photon : baryon ratio mentioned earlier. In the present universe, almost all the photons are in the cosmic background radiation – the number of photons that make up the radiation from stars and galaxies is negligible in comparison. Assuming the background radiation to have temperature 2.7 K, (8.11) then gives about 400 photons cm^{-3} as the present number density. We have seen earlier that there is an uncertainty in the present matter density of the universe. Assuming H_0 to be 50 km s^{-1} Mpc^{-1}, (4.9) gives 4.9×10^{-30} g cm^{-3} as the critical density. Let us suppose that the actual density is anywhere from 0.1 to 2 times the critical density. Since the matter is predominantly in baryons, this makes the baryon number density lie approximately between 0.3×10^{-6} and 6×10^{-6} per cubic centimetre (using the fact that a proton has mass 1.67×10^{-24} g). This implies that the ratio of baryons to photons lies approximately between 0.75×10^{-9} and 1.5×10^{-8}. Taking reciprocals, the ratio of photons to baryons is between 1.33×10^9 and 6.6×10^7. Although there is some uncertainty, the cosmic background radiation thus provides us with the useful piece of information of the approximate ratio of the numbers of photons and baryons. This number does not change as the universe evolves, unless it has gone through a stage which produces significant numbers of

photons through friction and viscosity, which seems unlikely if the standard model is correct. A knowledge of the photon:baryon number ratio enables us to infer the rate at which nucleosynthesis proceeded in the early universe, and to compare these predictions with the existing abundances of the nuclei. Although there are uncertainties in various stages, the above considerations do provide information about the different pieces in the jigsaw.

There is another way to look at the increase of wavelength of the background photons as the universe expands. Let R change by a factor f. Then the wavelength of a typical ray of light will also change by a factor f. This is clear from (3.52). After the expansion by a factor f the energy density $\mathrm{d}u'$ in the new wavelength range λ' and $\lambda' + \mathrm{d}\lambda'$ is decreased from the original energy density $\mathrm{d}u$ due to two effects: (*a*) since the number of photons in a given volume that has increased due to the expansion of the universe remains the same, the photon density decreases by a factor f^3; (*b*) since the energy of a photon is inversely proportional to its wavelength, its energy decreases by a factor f. Thus we get:

$$\begin{aligned}\mathrm{d}u' &= (1/f^4)\mathrm{d}u = (8\pi hc/\lambda^5 f^4)\mathrm{d}\lambda[\exp(hc/kT\lambda) - 1]^{-1}, \\ &= (8\pi hc/\lambda'^5)\mathrm{d}\lambda'[\exp(hcf/kT\lambda') - 1]^{-1}. \end{aligned} \tag{8.13}$$

This equation has the same form as (8.4) except that T has been replaced by T/f. It thus follows that freely expanding black-body radiation continues to be described by the Planck formula, but the temperature decreases in inverse proportion to R.

We can determine the neutrino temperature by considering the change in entropy as the universe expands. The entropy S at temperature T is proportional to $N_T T^3$, to a good approximation, where N_T is the effective number of species of particles in thermal equilibrium with threshold temperature below T. We have $N_T = N_1 N_2 N_3$, where N_1 is 1 if the particle does not have a distinct antiparticle, and 2 if it does; N_2 is the number of spin states of the particle; N_3 is a statistical mechanical factor which is $\frac{7}{8}$ or 1 according as to whether the particle is a fermion or a boson. In order to keep the total entropy constant, S must be proportional to R^{-3}, so that we have

$$N_T T^3 R^3 = \text{constant}. \tag{8.14}$$

As mentioned earlier, the neutrinos and antineutrinos went out of equilibrium with the rest of the contents of the universe before the annihilation of electrons and positrons (which occurred at approximately 5×10^9 K). Now according to the definition of N_T given above, electrons and

positrons have $N_T = \frac{7}{2}$, whereas photons have $N_T = 2$. Thus the total effective number of particles before and after the annihilation was

$$N_b = \tfrac{7}{2} + 2 = \tfrac{11}{2};\ N_a = 2. \tag{8.15}$$

From (8.14) it then follows that

$$\tfrac{11}{2}(T'R')^3 = 2(T''R'')^3, \tag{8.16}$$

where T', R' are values of T, R before annihilation, and T'', R'' the values afterwards. Thus

$$T''R''/T'R' = (\tfrac{11}{4})^{1/3} = 1.401. \tag{8.17}$$

This gives the increase in TR due to the heat produced by the annihilation. The neutrino temperature T'_ν before the annihilation was the same as the photon temperature T'; from then on T'_ν just decreased like R^{-1}. Let the neutrino temperature afterwards be T''_ν. Thus

$$T'_\nu R'' = T'_\nu R' = T'R', \tag{8.18}$$

from which, with the use of (7.17), it follows that

$$T''/T''_\nu = T''R''/T''_\nu R'' = T''R''/T'_\nu R' = T''R''/T'R' = 1.401. \tag{8.19}$$

Although the neutrinos go out of equilibrium quite early, they continue to make a significant contribution to the energy density. Remembering that the effective number of species N_T for neutrinos is $\frac{7}{2}$, and that the energy density is proportional to the fourth power of the temperature, the ratio of the densities of neutrinos to photons is:

$$u_\nu/u_\gamma = \tfrac{7}{4}(\tfrac{4}{11})^{4/3} = 0.4542. \tag{8.20}$$

From (8.9) we see that the photon energy density u_γ can be written as follows:

$$u_\gamma = 7.5641 \times 10^{-15} T^4 \text{ erg cm}^{-3}. \tag{8.21}$$

Thus the total energy density after the electrons and positrons have annihilated is

$$u = u_\gamma + u_\nu = 1.4542 u_\gamma = 1.100 \times 10^{-14} T^4 \text{ erg cm}^{-3}. \tag{8.22}$$

The equivalent mass density is as follows:

$$\text{mass density} = u/c^2 = 1.22 \times 10^{-35} T^4 \text{ g cm}^{-3}. \tag{8.23}$$

Given that the present temperature of the background radiation is of the order of 3 K, we see from (8.23) that the mass-energy density of this radia-

tion is negligible in comparison to that of visible matter, which is of the order of 10^{-31} g cm^{-3}.

We have said earlier that the temperature decreases as R^{-1}. To examine this further consider the situation in which the rest masses of the particles are not necessarily negligible in comparison with their kinetic energies. Then the mass-energy density and the pressure are given as follows (we revert to ε):

$$\varepsilon = mn + \tfrac{3}{2}nkT + N'aT^4, \tag{8.24a}$$

$$p = nkT + \tfrac{1}{3}N'aT^4, \tag{8.24b}$$

where we envisage the contents to have a common temperature T, m being the mass of the massive particles (nucleons), k, a are the Boltzmann and Stefan constants, n is the number density of nucleons, and N' is related to the number of species of particles. The first terms in (8.24a) and (8.24b) give the non-relativistic contributions, the later ones give the relativistic terms. The number density n satisfies the following equation:

$$n(t)R^3(t) = \text{constant}. \tag{8.25}$$

This can be established from the baryon conservation law

$$J^{\mu}{}_{;\mu} = 0, \tag{8.26}$$

where the baryon current J^{μ} is given by $J^{\mu} = nu^{\mu}$, u^{μ} being the four-velocity. Equation (8.25) is then obtained from (8.26) with the use of (2.6a), (3.72a) and (3.72b). We now substitute from (8.24a), (8.24b) and (8.25), into (3.79), which we write here again for convenience:

$$\dot{\varepsilon} + 3(p + \varepsilon)\dot{R}/R = 0. \tag{8.27}$$

The result of the substitution for $\dot{\varepsilon}$, $\dot{n}$, ε, p, into (8.27), is, after simplification, the following equation:

$$(\tfrac{1}{2} + N'\sigma)\dot{T}/T + (1 + N'\sigma)\dot{R}/R = 0, \tag{8.28}$$

where $\sigma = 4aT^3/3nk$. When $\sigma \gg 1$, (8.28) yields $TR = \text{constant}$ as a solution. In this case σ becomes a constant, since n varies as R^{-3}. This is termed a *hot universe*. To see what this implies, recall the number density of photons given by (8.11), which can be written as follows:

$$N = 20.28T^3 \text{ photons cm}^{-3} = 0.37(a/k)T^3 \text{ photons cm}^{-3}, \tag{8.29}$$

using the fact that $a = \pi^2k^4/15c^3\hbar^3 = 7.5641 \times 10^{-15}$ erg cm^{-3} K^{-4}, and $k = 1.38 \times 10^{-16}$ erg K^{-1}. Here $\hbar = h/2\pi$. From (8.29) and the definition of σ we see that

$$\sigma = 3.6\, N/n. \tag{8.30}$$

Thus the condition $\sigma \gg 1$ implies that there are very many more photons and other relativistic particles than nucleons, so that radiation is unaffected by matter and after the decoupling of matter and radiation the temperature continues to drop like R^{-1}. The radiation maintains its blackbody spectrum throughout the early universe as well as after the decoupling. We see from (8.25) and the decrease of T as R^{-1} that if the present number density of nucleons and the present temperature are respectively n_0 and T_0 and if these quantities have values n_1, T_1 respectively (the temperature being that of the background radiation; see the following sentence for a possible epoch to which n_1, T_1 refers), then the following relation obtains:

$$T_0 = (n_0/n_1)^{1/3} T_1. \tag{8.31}$$

If one can make a reasonable estimate of the nucleon number density at some early epoch, say when deuterium was just being formed (just below 10^9 K or so), one could predict the present temperature of the radiation from (8.31) and a knowledge of the present number density of nucleons. We will come back to this point later. Alternatively, one can use the present observed value 2.7 K of T_0 and an estimate of the present number density of nucleons to calculate the relation between T and n at any early epoch, and see what this implies for the abundances of the various nuclei. It is one of the successes of the standard model that the predictions of the abundances turn out to be in reasonable agreement with observed estimates.

8.5 Evolution of the mass-energy density

If we assume the early universe to be dominated by radiation, the equation of state is $p = \frac{1}{3}\varepsilon$, and (8.27) gives

$$\dot{R}/R = -\tfrac{1}{4}\dot{\varepsilon}/\varepsilon, \tag{8.32}$$

so that, with the use of (8.1) we get

$$\dot{\varepsilon} = -4(8\pi G/3)^{1/2}\varepsilon^{3/2}/c, \tag{8.33}$$

which can be integrated to give the following equation:

$$t = (3/32\pi G)^{1/2}\varepsilon^{-1/2}c + \text{constant}. \tag{8.34}$$

This relation, together with considerations of the previous section, leads to a thermal history of the early universe. This is done as follows. For any given range of temperatures, one determines the types of particles that are present in thermal equilibrium. One then determines the corresponding

mass-energy density, assuming the particles to be relativistic. The temperature is given by Stefan's T^4 law. One then gets a relation between the time and the temperature with the use of (8.34). We will follow this procedure to provide a more quantitative description of the evolution of the early universe than that given at the beginning of this chapter. In this we follow mainly the accounts given by Weinberg (1972) and Bose (1980). We may repeat some parts of the qualitative account given earlier.

(i) $10^{12}\ \mathrm{K} > T > 5.5 \times 10^9\ \mathrm{K}$

Just below 10^{12} K the matter in the early universe consists of photons (γ), electron–positron pairs (e^-, e^+), electron- and muon-neutrinos and their antiparticles (ν_e, ν_μ, $\bar{\nu}_e$, $\bar{\nu}_\mu$). There is also a small admixture of nucleons (neutrons and protons) and electrons – these will form the atoms of the later universe. Certain numbers of muons are also present to keep the neutrinos in thermal contact with other particles via weak interaction processes. The particles have a common temperature which is falling like R^{-1}. When the temperature goes below 10^{11} K or so, the neutrinos cease to be in thermal contact with the rest of the matter and radiation, but they continue to share a common temperature which drops like R^{-1}.

If the mixture of relativistic matter and radiation is considered to be an ideal gas, the number density $n_i(q)\,\mathrm{d}q$ of particles of species i with momentum between q and $q+\mathrm{d}q$ in thermal equilibrium is given as follows (Weinberg, 1972, Equation (15.6.3)):

$$n_i(q)\,\mathrm{d}q = (4\pi/h^3) g_i q^2 \{\exp[(E_i(q) - \mu_i)/kT] \pm 1\}\,\mathrm{d}q, \tag{8.35}$$

where the positive sign applies for fermions and the negative for bosons. Since the particles are relativistic, the energy $E_i(q)$ of the ith particles with mass m_i is given by $c(q^2 + c^2 m_i^2)^{1/2}$, μ_i is the chemical potential of the ith species, g_i is the number of spin states, with $g=1$ for neutrinos and antineutrinos, and $g=2$ for photons, electrons, muons and their antiparticles.

The energy density for the ith species is given by

$$\varepsilon_i = \int_0^\infty E_i(q) n_i(q)\,\mathrm{d}q, \tag{8.36}$$

so that with the use of (8.35) one gets the following values for the photon and neutrino densities:

$$\varepsilon_\gamma = aT^4;\ \varepsilon_\nu = \tfrac{7}{16} aT^4, \tag{8.37}$$

where a is Stefan's constant mentioned earlier. The chemical potential for the photon is zero, so that for electrons and positrons it is equal and

opposite, since chemical potential is conserved additively in reactions and $e^{\pm}$ pairs are produced from photons. However, in the range of temperatures under consideration, there are many more electron–positron pairs than unpaired electrons. Thus the number density of electrons is almost equal to that of positrons; since the corresponding chemical potentials are opposite it is reasonable to assume from (8.35) that both these chemical potentials vanish. Since the electrons are highly relativistic in the range of temperatures under consideration we can set $m_e \simeq 0$, and (8.36) yields the electron energy to be as follows:

$$\varepsilon_{e^-} = \tfrac{7}{8} a T^4. \tag{8.38}$$

One can use these parameters to calculate the electron number density from (8.35) as follows:

$$n_{e^-} = \tfrac{3}{4} N, \tag{8.39}$$

where N is the photon number density given by (8.29). Since n, the nucleon density, is nearly equal to the density of 'atomic' (unpaired) electrons, we see from (8.29), (8.30), (8.39) and the fact that $\sigma \gg 1$, that the electrons are predominantly the pair-produced ones. Adding the contributions of γ, ν_e, ν_μ, $\bar{\nu}_e$, $\bar{\nu}_\mu$, e^-, e^+, we get the total energy density to be as follows:

$$\varepsilon = \tfrac{9}{2} a T^4. \tag{8.40}$$

Putting this value of ε in (8.34) and inserting the values of a and G we get

$$t = 3.27 \times 10^{10}/T^2 + \text{constant} = 1.09/T'^2 \text{ (s)} + \text{constant}, \tag{8.41}$$

where T' is the temperature measured in units of 10^{10} K. Thus the temperature takes 0.0108 s to drop from $T' = 10^2$ (that is, $T = 10^{12}$ K) to $T' = 10$ K ($T = 10^{11}$ K) and another 1.079 s to drop to $T' \simeq 1$, ($T = 10^{10}$ K). These values are roughly consistent with the 'first frame' time and temperature $t = 0.01$ s, $T = 10^{11}$ K, and 'third frame' $t = 1.1$ s, $T = 10^{10}$ K.

(ii) $5.5 \times 10^9 \text{ K} > T > 10^9 \text{ K}$

We have $m_e = 0.51$ MeV, so that the rest mass of an electron–positron pair is about 1.02 MeV. Thus the temperature at which electron–positron pairs are produced is given by $kT \simeq 1.02$ MeV, which yields, using the fact that $k^{-1} = 11\,605 \text{ K eV}^{-1}$, a value of 1.1837×10^{10} K for the temperature at which pair production occurs. Thus at about 10^{10} K the electron–positron pairs start annihilating, and at the beginning of the present era these pairs become non-relativistic, so that (8.38) is no longer valid, and the behaviour $T \propto R^{-1}$ has to be modified. One can proceed by considering the entropy of

particles in thermal equilibrium: electrons, positrons and photons. With the use of (8.35) one can work out the entropy in a volume R^3, as follows:

$$S=\frac{R^3}{T}(\varepsilon+p)=\frac{4a}{3}(RT)^3\left\{1+\frac{15}{2\pi^4}\int_0^\infty \frac{x^2(x^2+3y^2)}{y[\exp(y)+1]}\,\mathrm{d}x\right\}, \tag{8.42}$$

where $x=q/kT$, $y=E/kT$, $E=c(q^2+c^2m_e^2)^{1/2}$. Since the entropy S is constant, one can use (8.42) to determine how T changes with R. When the electrons are relativistic, we have $m_e\approx 0$, $x\approx y$, and the expression in the curly brackets becomes $(1+\frac{7}{4})$; for non-relativistic electrons this factor is 1, so that $(RT)^3$ increases by a factor $\frac{11}{4}$. The ratio of the photon to neutrino temperatures, as we saw in (8.19), becomes $(\frac{11}{4})^{1/3}\approx 1.401$.

(iii) $T<10^9$ K

The electron–positron pairs have annihilated completely and the particles in equilibrium are photons and the relatively small number of 'atomic' electrons and nucleons. The neutrinos have been decoupled for some time and are expanding freely. The corresponding temperatures and energy densities are worked out in (8.17)–(8.23), with the electron–nucleon densities negligible at the beginning of this era. From (8.20)–(8.22) we see that the energy density in the early stages of this era is

$$\varepsilon=[1+\tfrac{7}{4}(\tfrac{4}{11})^{4/3}]aT^4=1.45aT^4. \tag{8.43}$$

Substituting in (7.34) we get

$$t=(15.5\pi Ga)^{-1/2}T^{-2}c+\text{constant}=192T''^{-2}\,(s)+\text{constant}, \tag{8.44}$$

where T'' is measured in units of 10^9 K. By putting T'' equal to 1 and 0.1 respectively and subtracting, we see that it took about 5 h and 16.8 min for the temperature to drop from 10^9 K to 10^8 K. Equation (8.44) also gives the age of the universe at the time of recombination, that is, when electrons and protons combined to form hydrogen atoms at a temperature of about 4000 K, of about 4×10^5 years.

The onset of the matter-dominated era can be worked out as as follows. From (8.25) and the dropping of the photon temperature as R^{-1} we see the number density of nucleons satisfies

$$n/n_0=(T/T_0)^3, \tag{8.45}$$

where n_0, T_0 are the present values of n, T. Thus the mass density of nucleons equals the density of radiation given by (8.43) at a temperature T_c which is as follows:

$$T_c=mn_0/(1.45aT_0^3). \tag{8.46}$$

If we take the present density of matter as 5×10^{-31} (this amounts to one-tenth of the critical density given by (4.9) if $H_0\simeq50$), then we get

$$T_c\sim2085\text{ K}. \tag{8.47}$$

and the corresponding age of the universe from (8.44) is approximately 1.6×10^6 years. Thus the ages at which matter started becoming dominant and at which recombination occurred are of the same order of magnitude.

Thus in the early universe particles were highly relativistic most of the time and (8.32) and (8.34) are valid for that period; ε is found to be proportional to T^4 and (8.41) and (8.44) are obtained as the time–temperature relations, so that T decreases as R^{-1}. However, during the brief period of electron–positron annihilation the more complicated relation (8.42) obtains, which can be written as

$$(RT)^3F(T)=\text{constant}, \tag{8.48}$$

where $F(T)$ is a complicated function which becomes a constant both in the highly relativistic and fully non-relativistic regimes, yielding the usual behaviour $RT=$constant, but with different constants in the two regimes. The function $F(T)$ can be worked out by numerical methods; this becomes necessary if one wants to follow the details of the temperature drop which may be required for an analysis of nucleosynthesis.

Before we end this section we show explicitly how some of the figures mentioned at the beginning of this chapter are arrived at from the formalism given in this and the last two sections. The time and temperature for the first and third frames have already been mentioned in the paragraph containing (8.41). We are only concerned with the approximate derivation of the figures; a precise number containing several significant figures is not very meaningful in view of the uncertainties mentioned earlier, such as the photon:baryon ratio.

If we assume that $t\ll0.01$ s for some large value of T such as 10^{14} K, we can take the constant in (8.41) as negligibly small for our purpose. Then if we set $T'=0.3$ K (which is the fourth frame temperature), we get $t=109/9\text{ s}\simeq12.1$ s. This is consistent with $t=13$ s mentioned for the fourth frame, because we are just outside the range for which (8.41) is applicable, and t is a little higher than that given by (8.41). For the fifth frame (8.44) is just beginning to be applicable and for this frame we have $T''=1$, so that (again assuming the constant to be negligible), $t=192$ s, which is consistent with t approximately 3 min given for the fifth frame. Similarly, for the sixth frame we put $T''=0.3$ K in (8.44) and

get approximately 35 min for t. Also, when T is 4000 K at recombination, we get t as several hundred thousand years from (8.44), as mentioned towards the end of Section 8.1.

As an example of the energy density, from (8.23), taking T to be the third frame temperature of 10^{10} K, we get ε to be 1.22×10^{5} cm^{-3}, which is consistent with the value mentioned for the third frame. Lastly, we give an example of the calculation of the Hubble time, which is the characteristic expansion time of the universe, given by $H^{-1}=R/\dot{R}$. From (8.32) and (8.33) we see that $R/\dot{R}=(8\pi G/3)^{-1/2}\varepsilon^{-1/2}$, which gives about 3 or 4 s as the third frame Hubble time, as mentioned.

8.6 Nucleosynthesis in the early universe

We have seen that in the early universe when the temperature was high enough neutrons and protons were separate and independent entities. In the present universe there are scarcely any free neutrons left; they form part of helium or heavier nuclei. In fact about 70–80% of the matter in the present universe is in the form of hydrogen, about 20–30% in the form of helium and a small percentage in the form of heavier nuclei. Any satisfactory theory of the early universe must explain the present observed abundances of the elements. One place in which nucleosynthesis can take place in the later universe, as mentioned earlier, is the centre of stars, where the temperature is of the order of a million Kelvin. Many of the heavier nuclei can indeed be produced here, as was shown in a famous paper by Burbidge, Burbidge, Fowler and Hoyle (1957). However, a simple calculation shows that the 20–30% helium that is observed today could not have been produced in the centre of stars. Indeed, the rate of energy release of our galaxy, for example, is about 0.2 erg g^{-1} s^{-1}. If the galaxy has been in existence for about 10^{10} years, this gives a total energy radiation of about 0.6×10^{17} erg per gram, or 0.375×10^{23} MeV per gram. Using the fact that a nucleon has mass 1.67×10^{-24} g, we see that this amounts to energy release of about 0.0625 MeV per nucleon, whereas hydrogen fusion into helium releases about 6 MeV per nucleon, so that only about 1% of the hydrogen in our galaxy could have been converted into helium.

In this section we will give an account of nucleosynthesis in the early universe. This is mainly based on Peebles (1971), Weinberg (1972), Schramm and Wagoner (1974) and Bose (1980).

The original suggestion that helium was synthesized in the early universe was made by Gamow, who developed a theory of nucleosynthesis

with his collaborators in the 1940s. Although this theory was incomplete in some respects, there were useful insights and, in fact, a cosmic background radiation with temperature of 5 K was predicted in the 1940s! However, for various reasons this theory was not taken seriously. Gamow realized that helium synthesis was possible only during a brief period in the early universe (the first few minutes) and that for a sufficient amount of helium to be produced the density must have been very high. This leads to the picture of a hot and dense early universe, a picture which is essential in understanding nucleosynthesis in the early universe. One can start with the presently observed 2.7 K as the temperature of the remnant radiation and work backwards. This was done by Peebles (1971) and with other reasonable assumptions he obtained a helium abundance of about 25%. This is one of the conspicuous successes of the picture of an early universe that is hot and dense.

To work out the details one has to determine how the neutron–proton balance changes as the universe evolves; see if the rate of deuterium formation is sufficiently fast to ensure that nearly all the neutrons are used up; and see if the reactions are fast enough to convert nearly all the deuterium into helium.

Neutrons and protons are converted into each other by the following weak processes:

$$n \leftrightarrow p + e^- + \bar{\nu}; \quad \mathrm{n} + \mathrm{e}^+ \leftrightarrow \mathrm{p} + \bar{\nu}; \quad \mathrm{n} + \nu \leftrightarrow \mathrm{p} + \mathrm{e}^-. \tag{8.49}$$

In the equilibrium condition as many neutrons are changing into protons as protons into neutrons. In the temperature range of interest the distribution of nucleons is given as follows, assuming they are non-relativistic (henceforth in the book we set $c = 1$ except for some specific cases):

$$n(q)\,\mathrm{d}q = (8\pi/h^3)\exp[(\mu - m)/kT - q^2/2mkT]q^2\,\mathrm{d}q. \tag{8.50}$$

Here μ is the chemical potential of neutrons and protons, these being the same since the chemical potential is additively conserved, as noted earlier, and since leptons have zero chemical potential. In (8.50) m is the mass of the nucleon in units of energy, with $m_n - m_p \equiv Q = 1.293$ MeV. Integrating (8.50) between zero and infinity and taking the ratio of the cases for neutrons and protons respectively, we get:

$$n'/n'' = \exp(-Q/kT), \tag{8.51}$$

where n' and n'' denote the neutron and proton number densities respectively. Note that n', n'' become equal as T tends to infinity, or t tends to zero.

The number densities for ν, $\bar{\nu}$, e^- and e^+ are given by (8.35) with zero chemical potential, with temperature T for $e^{\pm}$ and γ, and T_ν for ν, $\bar{\nu}$:

$$n_{e^-}(q)\,dq = n_{e^+}(q)\,dq = (8\pi/h^3)q^2\,dq\{\exp[E_e(q)/kT]+1\}^{-1}, \quad (8.52a)$$

$$n_\nu(q)\,dq = n_{\bar{\nu}}(q)\,dq = (4\pi/h^3)q^2\,dq\{\exp[E_\nu(q)/kT_\nu]+1\}^{-1}, \quad (8.52b)$$

where $E_e(q) = (q^2 + m_e^2)^{1/2}$ and $E_\nu(q) = q$, are the electron (or positron) and neutrino energies respectively. The rates of the reactions given in (8.49) are given by the V–A theory of weak interactions (see, for example, Marshak, Riazuddin and Ryan, 1969), with the proviso that the Pauli exclusion principle decreases these rates by a factor corresponding to fraction of states unfilled, as follows:

$$1 - [\exp(E_e/kT)+1]^{-1} = [1+\exp(-E_e/kT)]^{-1}, \quad (8.53a)$$

$$1 - [\exp(E_\nu/kT_\nu)+1]^{-1} = [1+\exp(-E_\nu/kT_\nu)]^{-1}. \quad (8.53b)$$

Taking into account (8.52a), (8.52b), (8.53a) and (8.53b), the rates of the processes (8.49) per nucleon are given as follows:

$$\lambda(n + \nu \to p + e^-) = A\int \upsilon_e E_e^2 q_\nu^2\,dq_\nu[\exp(E_\nu/kT_\nu)+1]^{-1}[1+\exp(-E_e/kT)]^{-1}, \quad (8.54a)$$

$$\lambda(n + e^+ \to p + \bar{\nu}) = A\int E_\nu^2 q_e^2\,dq_e[\exp(E_e/kT)+1]^{-1}[1+\exp(-E_\nu/kT_\nu)]^{-1}, \quad (8.54b)$$

$$\lambda(n \to p + e^- + \bar{\nu}) = A\int \upsilon_e E_\nu^2 E_e^2\,dq_\nu[1+\exp(-E_\nu/kT_\nu)]^{-1}[1+\exp(-E_e/kT)]^{-1}, \quad (8.54c)$$

$$\lambda(p + e^- \to n + \nu) = A\int E_\nu^2 q_e^2\,dq_e[\exp(E_e/kT)+1]^{-1}[1+\exp(-E_\nu/kT_\nu)]^{-1}, \quad (8.54d)$$

$$\lambda(p + \bar{\nu} \to n + e^+) = A\int \upsilon_e E_e^2 q_\nu^2\,dq_\nu[\exp(E_\nu/kT_\nu)+1]^{-1}[1+\exp(-E_e/kT_\nu)]^{-1}, \quad (8.54e)$$

$$\lambda(p + e^- + \bar{\nu} \to n) = A\int \upsilon_e E_e^2 q_\nu^2\,dq_\nu[\exp(E_e/kT)+1]^{-1}[\exp(E_\nu/kT_\nu)+1]^{-1}. \quad (8.54f)$$

The constant A here is given as follows:

$$A = (g_V^2 + 3g_A^2)/2\pi^3\hbar^7, \tag{8.55}$$

with g_V and g_A being the vector and axial vector coupling constants of the nucleon, with the following values:

$$g_V = 1.407 \times 10^{-49} \text{ erg cm}^{3;}\ g_A = -1.25 g_v, \tag{8.56}$$

which correspond to a half-life of $\sim$11 min for the decay of a free neutron. The lepton energies are related to Q as follows:

$$E_\nu + E_e = Q \quad \text{for} \quad n \leftrightarrow p + e^- + \bar{\nu}, \tag{8.57a}$$

$$E_\nu - E_e = Q \quad \text{for} \quad n + e^+ \leftrightarrow p + \bar{\nu}, \tag{8.57b}$$

$$E_e - E_\nu = Q \quad \text{for} \quad n + \nu \leftrightarrow p + e^-. \tag{8.57c}$$

In (8.54a)–(8.54f) v_e is the velocity of the electron given by q_e/E_e. These integrals are over lepton momenta that are consistent with (8.57a)–(8.57c). If these integrals are written over a common variable q ($=E_\nu = q_\nu$) in (8.54a) and (8.54d) and as $-E_\nu$ in (8.54b), (8.54c), (8.54e) and (8.54f) and we also replace $q_e^2\,\mathrm{d}q_e$ with $v_e E_e^2\,\mathrm{d}E_e$, the total transition rates for $n \to p$ and $p \to n$ can be written as follows:

$$\begin{aligned}\lambda(n \to p) &= \lambda(n \to p + e^- + \bar{\nu}) + \lambda(n + e^+ \to p + \bar{\nu}) + \lambda(n + \nu \to p + e^-) \\ &= A \int \left[1 - \frac{m_e^2}{(q+Q)^2}\right]^{1/2} (q+Q)^2 q^2\,\mathrm{d}q [1 + \exp(q/kT_\nu)]^{-1} \\ &\quad \times \{1 + \exp[-(q+Q)/kT]\}^{-1},\end{aligned} \tag{8.58a}$$

$$\begin{aligned}\lambda(p \to n) &= \lambda(p + e^- + \bar{\nu} \to n) + \lambda(p + \bar{\nu} \to n + e^+) + \lambda(p + e^- \to n + \nu) \\ &= A \int \left[1 - \frac{m_e^2}{(q+Q)^2}\right]^{1/2} (q+Q)^2 q^2\,\mathrm{d}q [1 + \exp(-q/kT_\nu)]^{-1} \\ &\quad \times \{1 + \exp[(q+Q)/kT]\}^{-1}.\end{aligned} \tag{8.58b}$$

Here T is the temperature of the electrons, photons and nucleons and T_ν is the neutrino temperature; below about 10^{10} K, T and T_ν are different and are given by (8.19). The integration in (8.58a) and (8.59b) ranges from $-\infty$ to $+\infty$ with a gap from $-Q - m_e$ to $-Q + m_e$. We are interested in the fractional abundance x given by

$$x = n'/(n' + n''), \tag{8.59}$$

whose evolution is given by the following equation:

$$\mathrm{d}x/\mathrm{d}t = -\lambda(n \to p)x + \lambda(p \to n)(1 - x). \tag{8.60}$$

Table 8.5 *Neutron fractional abundances as a function of time. (Taken from Peebles, 1971.)*

T(K)	t(s)	$\lambda(p\to n)$ (s^{-1})	$\lambda(n\to p)$ (s^{-1})	x
10^{12}	0.00010	4.02×10^{9}	4.08×10^{9}	0.496
10^{11}	0.0109	3.9×10^{4}	4.6×10^{4}	0.462
2×10^{10}	0.273	9	19	0.330
10^{10}	1.102	0.19	0.83	0.238
10^{9}	182	0	0.00109	0.130
8×10^{8}	296	0	0.00108	0.116
6×10^{8}	535	0	0.00107	0.089

In the limiting case when kT is much larger than Q, (8.58a) and (8.58b) yield the following approximations:

$$\lambda(p\to n)\simeq\lambda(n\to p)\simeq A\int q^4 dq[1+\exp(-q/kT)]^{-1}[1+\exp(q/kT)]^{-1},$$

$$=\tfrac{7}{15}\pi^4 A(kT)^5=0.36T'^5\ s^{-1}, \tag{8.61}$$

where, as in (8.41), T' is the temperature measured in units of 10^{10} K. We also have from (8.1) and (8.40):

$$\dot{R}/R=(12\pi aG)^{1/2}T^2=0.46T'^2\ s^{-1}. \tag{8.62}$$

From (8.61) and (8.62) we see that at $T'=1$ ($T=10^{10}$ K) a neutron is converting into a proton (and vice versa) at almost the same rate at which the universe is expanding. Thus at temperatures higher than 10^{10} K or so the processes (8.49) attain equilibrium and (8.50) is valid, and initially the neutron/proton numbers are nearly equal. Below 10^{10} K or so one has to integrate (8.58a), (8.58b), (8.59) and (8.60) numerically. This was done by Peebles (1971) and the results are set out in Table 8.5.

Helium synthesis involves essentially three steps. First, deuterium is produced (at a suitable temperature) directly from neutrons and protons. Next, two deuterium nuclei produce He^3 or H^3. The latter two nuclei then produce He^4, which is the stable helium isotope. The precise working out of helium synthesis is a complicated matter involving many equations. Such details have been considered by Peebles (1966) and by Wagoner, Fowler and Hoyle (1967). The reactions involved are many, such as:

$$\left.\begin{aligned} &\mathrm{p}+\mathrm{n}\leftrightarrow\mathrm{D}+\gamma;\ \mathrm{D}+\mathrm{D}\leftrightarrow\mathrm{He}^3+\mathrm{n}\leftrightarrow\mathrm{H}^3+\mathrm{p};\\ &\mathrm{H}^3+\mathrm{D}\leftrightarrow\mathrm{He}^4+\mathrm{n};\ \mathrm{p}+\mathrm{D}\leftrightarrow\mathrm{He}^3+\gamma;\ \mathrm{n}+\mathrm{D}\leftrightarrow\mathrm{H}^3+\gamma;\\ &\mathrm{p}+\mathrm{H}^3\leftrightarrow\mathrm{He}^4+\gamma;\ \mathrm{n}+\mathrm{He}^3\leftrightarrow\mathrm{He}^4+\gamma;\ \mathrm{D}+\mathrm{D}\leftrightarrow\mathrm{He}^4+\gamma. \end{aligned}\right\} \tag{8.63}$$

Reactions involving γs (photons) are radiative processes which usually take longer than other ones. Nucleosynthesis, when it begins, proceeds very quickly. The precise temperature at which it begins depends on the density, which can be extrapolated backwards from the present density, knowing the temperature of the background radiation. Peebles finds that nucleosynthesis begins at $T=0.9\times10^9$ K if the present density is $\varepsilon_0\simeq7\times10^{-31}$ g cm^{-3}, or at $T=1.1\times10^9$ K if it is $\varepsilon_0\simeq1.8\times10^{-29}$ g cm^{-3}. All processes which are relevant conserve the total number of nucleons. One result of nucleosynthesis is that the neutron:proton ratio is 'frozen' at the value it had just before nucleosynthesis began because once inside a nucleus a neutron cannot undergo beta decay. Before nucleosynthesis began, the ratio of neutrons to all nucleons is given by x (see (8.59)). After nucleosynthesis there are just free protons and He^4 nuclei. Thus the fraction of neutrons to all nucleons is just half the fraction of nucleons bound in He^4; this is the same as the abundance of helium by weight. It is found that a probable value (which comes out of the above calculations of x) when nucleosynthesis begins is 0.12. Thus the theory predicts about 24% for helium abundance, which is consistent with the observed value.

Appreciable amounts of elements heavier than helium cannot be produced in the early universe as there are no stable nuclei with five or eight nucleons, as mentioned earlier. As regards nuclei with seven nucleons, the Coulomb barrier (repulsion between the protons in different nuclei) in the reactions

$$\mathrm{He}^4+\mathrm{H}^3\rightarrow\mathrm{Li}^7+\gamma;\ \mathrm{He}^4+\mathrm{He}^3\rightarrow\mathrm{Be}^7+\gamma, \tag{8.64}$$

prevents these in comparison with

$$\mathrm{p}+\mathrm{H}^3\rightarrow\mathrm{He}^4+\gamma;\ \mathrm{n}+\mathrm{He}^3\rightarrow\mathrm{He}^4+\gamma. \tag{8.65}$$

He^4 has the highest binding energy by far of all nuclei with less than five nucleons, so effectively all the neutrons are used up in the formation of He^4.

There is a simpler way of obtaining the neutron:proton ratio by comparing the weak interaction rate with the Hubble rate (as pointed out by Barrow, 1993). However, some of the details of the derivation given here, although circuitous, may be useful in other contexts.

8.7 Further remarks about helium and deuterium

We have seen earlier that the standard model predicts that the proportion of helium and deuterium present in the universe depends on the baryon:photon ratio. The helium abundance is higher for a greater number of baryons, while the deuterium abundance is correspondingly lower. The baryon:photon ratio is thus a crucial parameter in cosmology. As the cosmic background temperature is known fairly accurately, and as the photons in the present universe reside predominantly in the background radiation, the baryon:photon ratio can be worked out if one knows the matter density of the present universe, as the matter is predominantly in the form of baryons. Thus an accurate observational determination of the matter density, and of the relative abundances of helium and deuterium, can provide a useful test of the standard model.

To settle this question one has to examine if there are processes in the later universe which can create or destroy helium and deuterium. As we remarked earlier, significant amounts of helium could not have been produced in the later universe. One has to ask a similar question about deuterium. A brief discussion of deuterium production and destruction is in order here. In the Sun and such typical hydrogen burning main sequence stars deuterium is produced by weak interaction as follows:

$$\mathrm{p} + \mathrm{p} \rightarrow \mathrm{D} + \mathrm{e}^{+} + \nu_{\mathrm{e}}. \tag{8.66}$$

The deuterium thus produced is quickly transformed by the much faster reaction

$$\mathrm{p} + \mathrm{D} \rightarrow \mathrm{He}^{3} + \gamma. \tag{8.67}$$

Reactions (8.66) and (8.67) lead to a small equilibrium abundance of deuterium. The small amount of deuterium that is present in the interstellar medium and that is incorporated in stars soon disintegrates due to reactions such as (8.67). Thus any deuterium that existed when the galaxy was formed would be depleted by now. As mentioned earlier, deuterium is also created by the following radiative process:

$$\mathrm{p} + \mathrm{n} \rightarrow \mathrm{D} + \gamma, \tag{8.68}$$

which is not prevented by the Coulomb barrier and involves no weak interaction. However, the free neutron that (8.68) requires is not usually present in astrophysical situations, except where there is very high energy involved such as in supernova explosions.

Another astrophysical situation in which deuterium can be created is in spallation reactions, mainly through the following reaction:

$$p + He^4 \rightarrow D + He^3. \tag{8.69}$$

This requires a centre-of-mass energy of 18.35 MeV, which is very high, because the binding energies of the product nuclei are somewhat less than those of the initial ones. In (8.69), for example, a part of this energy is used up in extracting the neutron from the He^4 nucleus. Such high energies sometimes exist in cosmic ray protons.

In astrophysical settings deuterium can be readily destroyed by the following reactions:

$$D + D \rightarrow n + He^3; \quad D + D \rightarrow H^3 + p; \quad n + D \rightarrow H^3 + \gamma, \tag{8.70}$$

if either the neutron or deuterium concentration is high. Thus to produce and preserve deuterium one needs energy and low density.

The abundance of deuterium is usually specified by D/H, the ratio of deuterium and hydrogen nuclei in a small volume. This ratio is different in different astrophysical and terrestrial situations. In sea water, for example, where deuterium occurs as HDO (heavy water; obtained by replacing a hydrogen atom in H_2O by deuterium), the ratio is 150 ppm (parts per million), which is somewhat higher than the average for all situations. The proportion of deuterium in carbonaceous meteorites is similarly high. The high proportion of deuterium in sea water is explained by the fact that due to chemical fractionation, in the formation of water D is preferred to H; the larger mass of D allows for different chemical and nuclear properties. On the other hand, in the outer regions of the Sun D/H is only about 4 ppm. This is because reactions such as $D + p \rightarrow He^3 + \gamma$, destroy deuterium in the Sun. In the interstellar gas near the Sun D/H is about 14 ppm. In the interstellar gas deuterium is detected through deuterated molecules such as CH_3D (deuterated methane) and DCN (deuterium cyanide). For example, it was found that in the Orion nebula the DCN/HCN ratio was about 40 times the terrestrial D/H ratio (Jefferts, Penzias and Wilson, 1973; Wilson, Penzias, Jefferts and Solomon, 1973); this is again due to chemical fractionation which favours DCN formation over HCN. The deuterium in interstellar material is detected by its 91.6 cm hyperfine line (the equivalent of the well-known 21 cm hydrogen line). The possibility of deuterium production in supernova explosions has also been considered (see Schramm and Wagoner, 1974, for references on this), but it is found that these explosions are much more efficient at producing other light elements such as Li^7, Be^9 and B^{11} than D. In Table 8.6 we set out the observed

Table 8.6 *Observed ratio of deuterium to hydrogen atoms. (Reproduced from Schramm and Wagoner, 1974, with minor omissions.)*

	Location	$(D/H)\times 10^6$ (ppm)	Observer
Solar system	Earth (HDO)	150	Friedman *et al.*, 1964
	Meteorites (HDO)	130–200	Boato, 1954
	Jupiter (CH_3D)	28–75	Beer and Taylor, 1973
	Jupiter (HD)	21 ± 4	Trauger *et al.*, 1973
	Present Sun	<4	Grevesse, 1970
	Primordial Sun:		
	From He^3 in gas-rich meteorites	10–30	Black, 1971, 1972
	From He^3 in solar wind	<50	Geiss and Reeves, 1972
	From He^3 in solar prominences	<60	Hall, unpublished
Interstellar medium	Cassiopeia A (91.6 cm line)	<70	Weinreb, 1962
	Sagittarius A (1.6 cm line)	<350	Cesarsky, Moffet and Pasachoff, 1973, Pasachoff and Cesarsky, 1974
	β Centauri	14 ± 2	Rogerson and York, 1973

abundances of deuterium in various situations; although some of these may be out of date the table nevertheless incorporates some essential points.

As the Jovian CH_3D estimate requires determination of the CH_4 abundance there was some uncertainty about this measurement (Beer and Taylor, 1973). The Voyager infrared experiment enabled a simultaneous determination of CH_4/CH_3D mixing ratio and Kunde *et al.* (1982) then derived the D/H ratio from Jovian CH_3D as 22 and 46 ppm (see Gautier and Owen (1983)). This is not inconsistent with the 1973 estimate given by Beer and Taylor (see Table 8.6).

The question arises as to what extent the helium and deuterium abundances found in the present universe represent these abundances in the primordial universe soon after nucleosynthesis. As we have seen, deuterium can be created to a small extent and destroyed more readily in the later universe. For helium a minor component of the abundance currently observed can be produced in stars and injected into the interstellar medium by supernova explosions and stellar winds. The giant planets like Jupiter and Saturn, because of their low exospheric temperatures and large masses, provide environments in which the elements are more or less in their primordial form, almost undisturbed for about 4.55 billion years since these planets were formed. Even the lightest elements have not escaped from the atmospheres of these planets since their inception. The Jovian helium abundance has been determined by Voyager. The hydrogen/helium mixing ratio can be found in many different areas of Jupiter and one finds a mass ratio Y (Y = mass of helium/mass of all nuclei) of 0.19 ± 0.05 by one method, and $Y = 0.21 \pm 0.06$ by another. Combining the two methods one gets (Gautier and Owen, 1983):

$$0.15 < Y < 0.24. \tag{8.71}$$

Table 8.7 summarizes a representative set of observations of helium abundance. The low value of Y for Saturn is probably due to the phenomenon of differentiation of helium from hydrogen (Smoluchowski, 1967), that may have depleted the amount of helium in the Saturn atmosphere. Presumably this phenomenon has not begun in Jupiter.

Gautier and Owen (1983) find that the primordial abundance of deuterium must have been reduced (that is, the deuterium must have been destroyed) by a factor of between 5 and 16 between the time of the primordial nucleosynthesis and the origin of the solar system 4.55 billion years ago. This is seen as follows. Since helium and deuterium were synthesized at the same time, Y and X(D) (the deuterium mass fraction which is

Table 8.7 *Helium abundances. (Taken from Gautier and Owen, 1983, with some omissions and a minor change.)*

Determination	Y	Reference
Jupiter (Voyager IRIS)	$0.15 < Y < 0.24$	Gautier *et al.*, 1981
Saturn:		
Pioneer 11	0.18 ± 0.05	Orton and Ingersoll, 1980
Voyager IRIS	~ 0.14	Conrath, Gautier and Hornstein, 1982
Solar:		
Helium emission lines	0.28 ± 0.05	Heasley and Milkey, 1978
Cosmic rays	0.20 ± 0.04	Lambert, 1967
Standard interior models	0.22	Iben, 1969; Bahcall *et al.*, 1973, 1980; Ulrich and Rood, 1973; Mazzitelli, 1979
Primordial		
best estimate from several results	0.23 ± 0.01	Pagel, 1984

approximately 1.5 times D/H – the exact multiple depending on Y) have a certain dependence on η, the ratio of baryon to photon number densities. The uncertainty in Y_p (the primordial value of Y) is found to be: $0.22 < Y_p < 0.24$, which corresponds to the following uncertainty in η: $0.35 \times 10^{-11} < \eta < 2 \times 10^{-10}$. The corresponding abundance of primordial deuterium turns out to be $3.4 \times 10^{-4} < (X(\mathrm{D}))_p < 11.6 \times 10^{-4}$. From the Jovian deuterium abundance one gets the upper limit $X(\mathrm{D}) < 7 \times 10^{-5}$. This leads to the discrepancy cited at the beginning of this paragraph. This analysis seems to imply that either deuterium is destroyed more efficiently than hitherto assumed, or that the standard model needs some modification. Whether this claim made by Gautier and Owen is valid is not clear, but the above analysis does emphasize the need to look very carefully into the question of helium and deuterium abundances and their relation with the baryon to photon number density ratio, both observationally and theoretically. There are some other assumptions made in this analysis which we have not mentioned; one of these is the assumption that there are three different kinds of neutrinos. The reader is referred to Gautier and Owen (1983) for more details.

As noted earlier, different amounts of nuclei are created in the early universe according to different assumptions of the baryon:photon ratio, which, in turn, depends on the present mass density of the universe. Thus

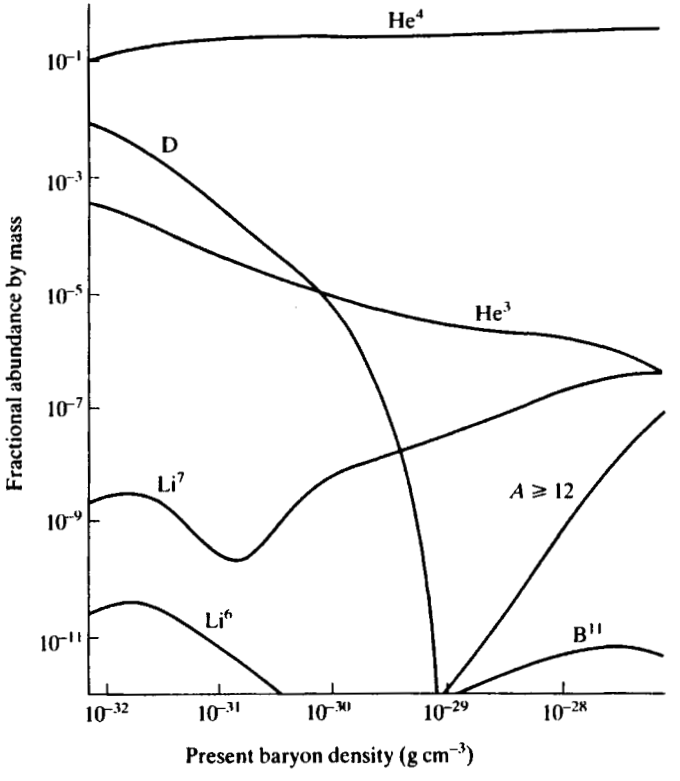


Fig. 8.3. This figure gives the dependence of the abundances of various nuclei on the present value of the mass density, which is not precisely known. The curve marked $A \geqslant 12$ refers to nuclei with baryon number greater than or equal to 12.

different values of the present mass density give different abundances. Figure 8.3 depicts this dependence of the abundances on the mass density, as given by Schramm and Wagoner (1974). It is interesting that the He^4 abundance is almost constant, that is, it is not at all sensitive to the value of the present mass density. By contrast, the abundance of D is strongly dependent on the mass density.

8.8 Neutrino types and masses

We end this chapter with a brief discussion of neutrino types and masses and the cosmological implications of these. We saw earlier that the temperature depends on the types of particles that were in thermal

equilibrium with the photons in the early universe. In our earlier analysis we did not adequately take into account the fact that there are different types of neutrinos. Two types, the electron- and muon-neutrinos, are definitely known. There may be a third type associated with the heavier tau-lepton, which was discovered relatively recently. If there are three or more kinds of neutrinos, it can be shown that this results in faster expansion in the early universe, so that more He^4 is produced. However, like the mass density, the He^4 abundance is not so sensitively dependent on neutrino types so that many more types than are known at present can be accommodated without seriously violating the observed He^4 abundance. However, it is quite a different matter with D abundance, which is highly sensitive to the number of neutrino types. It would be very difficult to reconcile the observed D abundance if the neutrino types were five or six in number. However, the latter situation would be saved somewhat if the neutrinos had mass, as has been indicated recently. Massive neutrinos have much less effect on the expansion rate and nucleosynthesis in the early universe. Another consequence of massive neutrinos is that the 'background' neutrinos then might contribute enough mass to the present mass density to make it above the critical density. The present indications are that neutrino masses cannot be more than a few electron volts. A recent analysis of the neutrino arrival time from the supernova in the Large Magellanic Cloud (Hirata *et al.*, 1987; Adams, 1988) shows that there is a 90% probability of the neutrino mass being less than 5 eV and 99% probability of it being less than 10 eV. A great deal of theoretical and observational work has to be done to clarify this question. We refer the interested reader to the papers cited, and Tayler (1983), Schramm (1982) and Bahcall and Haxton (1989). This question will be discussed further in the Appendix at the end of the book.

9

The very early universe and inflation

9.1 Introduction

As is clear from the discussion so far in this book, the standard big bang model incorporates three important observations about the universe. These are, firstly, the expansion of the universe discovered by Hubble in the 1930s, secondly, the discovery of the microwave background radiation by Penzias and Wilson and its confirmation by other observers and, thirdly, the prediction of the abundances of various nuclei on the basis of nucleosynthesis in the early universe, particularly the abundances of He^4 and deuterium, which appear to conform reasonably with observations. As is also clear from the earlier discussions, much theoretical and observational work remains to be done to clarify these questions further.

As mentioned in Chapter 1, some glaring puzzles do remain, such as the horizon problem. The puzzle here is: how is the universe so homogeneous and isotropic to such vast distances, extending to regions which could not have communicated with each other during the early eras? This problem is illustrated in Fig. 9.1. Another puzzle is why the density parameter Ω (the ratio of the energy density of the universe to the critical density – see discussion following (1.4)) is so near unity. If the present value of Ω lying between 0.1 and 2 is extrapolated to near the big bang we get the following orders of magnitude:

$$|\Omega(1\text{ s}) - 1| = O(10^{-16}), \tag{9.1a}$$

$$|\Omega(10^{-43}\text{ s}) - 1| = O(10^{-60}). \tag{9.1b}$$

These extremely small numbers seem difficult to explain. The third problem is the smoothness problem, which is to explain the origin and nature of the primordial density perturbations which result in the 'lumpiness', that is, the presence of galaxies and the structure of the observable

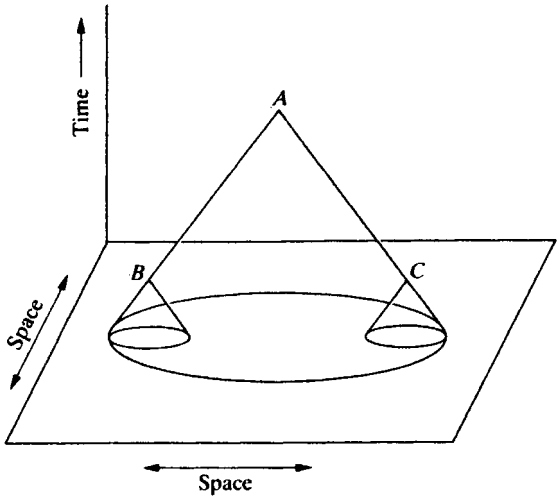


Fig. 9.1. This diagram illustrates the horizon problem. The point A represents our present space-time position, one space dimension being suppressed in this diagram. The points B and C represent events at a much earlier epoch, lying in opposite spatial direction from us, but lying in our past light cone. The plane at the bottom represents the instant $t=0$, the big bang. The past light cones of B and C have no intersection, so these two events could not have had any causal connection. How is it that radiation received from these two points (the cosmic background radiation) are at the same temperature?

universe. The inflationary models, of which the original one was propounded by Guth (1981), attempt to explain these puzzles.

9.2 Inflationary models – qualitative discussion

In this section we shall give a qualitative description of inflationary models; this will be followed by some quantitative accounts. However, it will not be possible to explain all aspects quantitatively. Some aspects involve fairly technical questions of particle physics and in particular Grand Unified Theories, which are beyond the scope of this book. Our treatment of inflationary models is by no means exhaustive; our intention is to point out the essential features.

As mentioned earlier, at high energies, according to the Glashow–Weinberg–Salam unified electroweak theory, electromagnetic and weak interactions behave in a similar manner and, consequently, there is a phase transition in the early universe associated with this at a critical temperature

of about 3×10^{15} K. The Grand Unified Theories attempt to find a unified description of all three of the fundamental interactions, namely, electromagnetic, weak and strong interactions. Grand Unified Theories predict that there is a phase transition in the universe at a critical temperature of about 10^{27} K, above which there was a symmetry among the three interactions. Consider again the analogy with the freezing of water. In the liquid state there is rotational symmetry at any point in the body of the water; this symmetry is lost, or 'broken' when ice is formed, as ice crystals have certain preferential directions. Secondly, the liquids in different portions begin to freeze independently of each other with different crystal axes, so that when the whole body of the liquid is frozen certain defects remain at the boundaries of the different portions. In a similar manner in the early universe above 10^{27} K or so the symmetry among the three interactions was manifest, and below this temperature this symmetry was broken. Now in water the manner in which the rotational symmetry is broken in different portions can be characterized by parameters which describe the orientation of the ice-crystal axes. Thus these parameters take different values in different portions of the liquid as it freezes, that is, as the symmetry is broken. In a similar way, the manner in which the manifest symmetry among the three interactions is broken can be characterized by the acquiring of certain non-zero values of parameters known as Higgs fields; this is referred to as *spontaneous symmetry breaking*. The symmetry is manifest when the Higgs fields have the value zero; it is spontaneously broken whenever at least one of the Higgs fields becomes non-zero. Just as in the case of the freezing of water, certain defects remain at the boundaries of different regions in which the symmetry is broken in different ways, that is, by the acquiring of different sets of values for the Higgs fields. There are point-like defects which correspond to magnetic monopoles, and two-dimensional defects called domain walls. A region in which the symmetry is broken in a particular manner could not have been significantly larger than the horizon distance at that time, so one can work out the minimum number of defects that must have occurred during the phase transition. The defects are expected to be very stable and massive. For example, it turns out that monopoles are about 10^{16} times as massive as a proton. The result is that there would be so many defects that the mass density would accelerate the subsequent evolution of the universe, so that the 3 K background radiation would be reached only a few tens of thousands of years after the big bang instead of ten billion years. Thus this prediction of Grand Unified Theories seems to conflict seriously with the standard model.

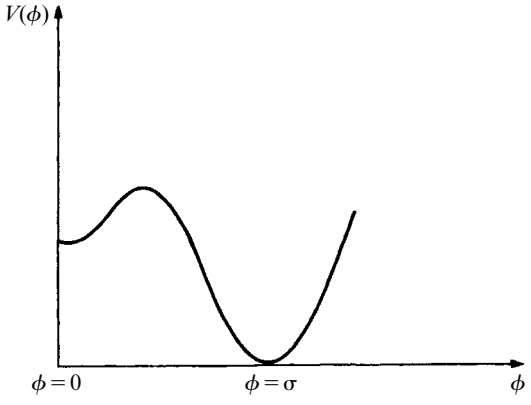


Fig. 9.2. One of the possible forms of the potential for the scalar field given by Equation (9.16a).

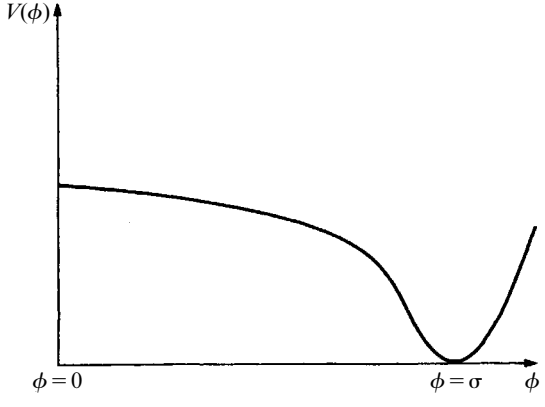


Fig. 9.3. Another possible form for the potential of the scalar field given by Equation (9.16b).

None of the successes of the standard model are affected by the inflationary models, because after the first 10^{-34} s or so, the two models are exactly the same as far as our observable universe is concerned. The original inflationary model put forward by Guth in 1981 had serious drawbacks, as mentioned in Chapter 1. We shall be concerned with the 'new inflation' put forward independently by Linde (1982) and Albrecht and Steinhardt (1982). For simplicity we consider a single Higgs field which we take to be a scalar field ϕ. The possible forms of the potential energy corresponding to this field are indicated in Figs. 9.2 and 9.3.

Consider some properties of the potential as depicted in Figs. 9.2 and

9.3. The potential has stationary points at $\phi = 0$ and $\phi = \sigma$. At these points the system can be in equilibrium. The states which are stationary states of the potential can be referred to as 'vacuum' states. Consider Fig. 9.2 first. The energy of the stationary state at $\phi = 0$ is higher than that at $\phi = \sigma$. There might be a situation in which the system is 'trapped' in the stationary state at $\phi = 0$ and cannot make the transition to the stationary state at $\phi = \sigma$, because of the potential barrier, even though $\phi = \sigma$ has a lower energy. In this situation the state $\phi = 0$ is referred to as a 'false vacuum', while $\phi = \sigma$ is the 'true vacuum'. What is the relevance of this to the very early universe and inflation?

We assume that the very early universe had regions that were hotter than 10^{27} K and were expanding. The symmetry among the interactions was manifest and the Higgs field, represented by ϕ here, was zero. One can look upon this situation as the thermal fluctuations driving the Higgs field to the equilibrium value zero. As the expansion caused the temperature to fall below the critical temperature, it would be thermodynamically more favourable for the Higgs field to acquire a non-zero value. However, for some values of the parameters in Grand Unified Theories the phase transition occurs very slowly compared to the cooling rate. This can cause the temperature to fall well below 10^{27}, the critical temperature, but the Higgs field to remain zero. This is akin to the phenomenon of supercooling; for example, water can be supercooled to 20° below freezing. This is the situation of the false vacuum mentioned above, in which the Higgs field remains zero although it is energetically more favourable to go to the state $\phi = \sigma$ (that is, the energy in the state $\phi = \sigma$ is lower than that in $\phi = 0$). It turns out that this situation causes the region to cool down considerably and also have a very high rate of expansion. The situation, depicted in Fig. 9.2, however, leads to difficulties, which are avoided in that depicted in Fig. 9.3, so we shall follow the rest of the development in the latter situation.

Before considering a more quantitative description of inflationary models, it may be useful to give an idea of the overall effect of these models on the standard model. This is given in Fig. 9.4, which is taken from Turner (1985). The inflationary models incorporate all the predictions of the standard model for the observable universe, because for the latter the inflationary models have the same behaviour after $t = 10^{-32}$ s or so. From about 10^{-34} to 10^{-32} s or so, the inflationary models are radically different. A region of the universe underwent accelerated exponential expansion, as well as cooling. After this period of expansion and cooling it was reheated to just below the critical temperature. After this the story is the same as the standard model, the important difference being, the initial

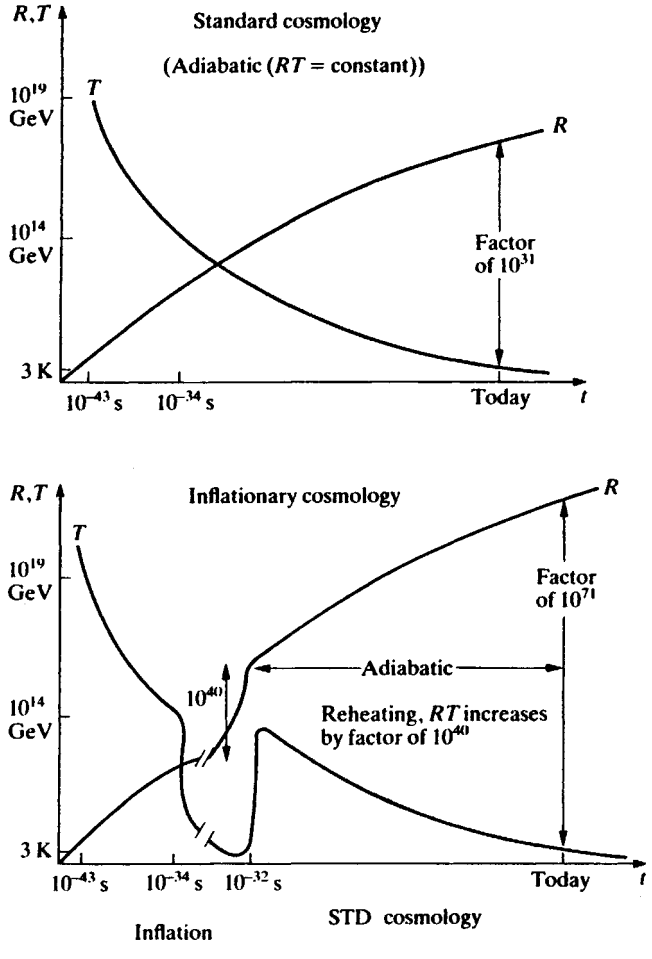


Fig. 9.4. This figure depicts the evolution of the scale factor R and temperature T of the universe in the standard model and in the inflationary models. The standard model is always adiabatic ($RT \approx$ constant), except for minor deviations when particle–antiparticle pairs annihilate, whereas inflationary models undergo a highly non-adiabatic event (at 10^{-34} s or so), after which it is adiabatic Turner, 1985).

region was within a horizon distance and had time to homogenize and have the same temperature, etc., and after the inflation the entire observable universe can lie within such a region, so that the horizon problem does not arise. Let us see this, still qualitatively, in more detail keeping in mind the Higgs potential of Fig. 9.3.

Consider a region of the very early universe which was hotter than 10^{27} K

or so. We will consider the evolution of this region in the inflationary model. The reason this evolution is different from the standard model is due to the presence of the Higgs field, which describes the phase transition that the universe undergoes at that temperature, as mentioned earlier. The presence of the Higgs field radically alters the evolution of the scale factor R in the very early epochs, as depicted in Fig. 9.4, the evolution of R being the same as in the standard model after 10^{-32} s or so. We will write down these equations in the next section. Here we describe this evolution qualitatively with the Higgs potential being given by Fig. 9.3. As mentioned earlier, above 10^{27} K thermal fluctuations drive the equilibrium value of the Higgs field to zero and the symmetry is manifest. As the temperature falls the system undergoes a phase transition with at least one of the Higgs fields acquiring a non-zero value (here we consider only one), resulting in a broken symmetry phase. However, for certain values of the parameters, which we assume to be the case, the rate of the phase transition is very slow compared with the rate of cooling. This causes the system to supercool to a negligible temperature with the Higgs field remaining at zero (this corresponds to the 'well' in the curve marked T in the lower figure in Fig. 9.4), resulting in a 'false vacuum'. Now quantum fluctuations or small residual thermal fluctuations cause the Higgs field to deviate from zero. Unlike the situation depicted in Fig. 9.2, in Fig. 9.3 there is no energy barrier, so the Higgs field begins to increase steadily. The rate of increase is, in fact, like the speed of a ball which was perched on top of the potential curve in Fig. 9.3 (at $\phi = 0$) and which starts to roll down; at first the speed is very slow, increasing gradually until it has high speed in the steeper portions, and finally it oscillates back and forth when it reaches the bottom of the well. In the flatter portions, as we shall see more clearly in the next section, the region undergoes accelerated expansion, doubling in diameter every 10^{-34} s or so. When the value of the Higgs field reaches the steeper parts of the potential curve, the expansion ceases to accelerate. An expansion factor of 10^{50} or more can be achieved in this manner for the region under consideration.

The picture given above is a simplified one. As mentioned earlier, there can be many different broken-symmetry states (depending on the non-zero values acquired by the Higgs fields), just as there are many different possible crystal axes during the freezing of a liquid. Thus different regions in the very early universe would acquire different broken-symmetry states, each region being roughly of the size of the horizon distance at the time. The horizon distance at time t is approximately ct, the distance travelled by light in time t; thus at $t = 10^{-34}$ s the horizon distance is about 10^{-24} cm.

Once a domain was formed with a particular set of non-zero values of the Higgs fields, it would gradually attain one of the stable broken-symmetry states and inflate by a factor of 10^{50} or so. Thus after inflation the size of such a domain would be approximately 10^{26} cm. At that epoch the entire observable universe would measure only 10 cm or so, so it would easily fit well within a single domain. Since the observable universe lay within a region which, in turn, started from a region contained in a horizon distance, it would have had time to homogenize and attain a uniform temperature. This then solves the horizon problem.

Because of the enormous inflation, any particle with a certain density that may have been present before the inflation, would have its density reduced to almost zero after the inflation. Most of the energy density would be incorporated in the Higgs field after the inflation. After the Higgs field evolves away from the flatter portion of the curve in Fig. 9.3 and goes down the steep slope and starts oscillating back and forth near the true vacuum at $\phi = \sigma$, we have the situation that corresponds in quantum field theory to a high density of Higgs particles (recall a high level of energy for a harmonic oscillator corresponds to a larger number of 'excitations' of the electromagnetic field, that is, a large number of photons). The Higgs particles would be unstable and would undergo decays into lighter particles, and the system would rapidly attain the condition of a hot gas of elementary particles in equilibrium, akin to the initial condition assumed in the standard model. The system would be reheated to a temperature of about 2–10 times lower than the phase transition temperature of 10^{27} K. The story after this is the same as the standard model, so that the successes of the standard model are maintained.

Several points and questions remain in the above description, which we will deal with at the end of this chapter. Firstly, how is the monopole problem solved by this model? One of the problems of the standard model that Grand Unified Theories purport to solve is the problem of baryon asymmetry, that is, secondly, why do we see matter rather than antimatter in the present universe? In other words, when in the last chapter we spoke of a small 'contamination' of neutrons and protons, one can ask why there was not a contamination of antineutrons and antiprotons instead. Thirdly, as a matter of interest, what was wrong with the original model put forward by Guth (1981)? Lastly, do any problems remain in the new inflationary model as described above? In other words, is the new inflationary model able to solve all the problems of the standard model mentioned earlier and not throw up problems of its own, that is, is it self-consistent and in accord with present observations?

9.3 Inflationary models – quantitative description

As mentioned earlier, it is not possible to give here the technical details from particle physics and quantum field theory. Secondly, even in the classical and semi-classical treatments, suitable exact solutions are not known so that even when we have the equations a certain amount of qualitative analysis is necessary.

Recall Einstein's equations (2.22) (with $c=1$):

$$R_{\mu\nu}-\tfrac{1}{2}g_{\mu\nu}R=T_{\mu\nu}.$$

Here $T_{\mu\nu}$ represents the energy–momentum tensor. For a perfect fluid this is given by (2.23). However, although the latter case suffices for the standard model, in general, one has to consider the contributions to $T_{\mu\nu}$ from all possible fields. For example, when there is an electromagnetic field present (this is not relevant in the cosmological context), one has to add the following contribution to the energy–momentum tensor:

$$T^{(\mathrm{em})}_{\mu\nu}=(4\pi)^{-1}(-F_{\mu}{}^{\alpha}F_{\nu\alpha}+\tfrac{1}{4}g_{\mu\nu}F_{\alpha\beta}F^{\alpha\beta}), \tag{9.2}$$

where the electromagnetic field tensor $F_{\mu\nu}$ is given in terms of the four-potential A_{μ} as follows: $F_{\mu\nu}=A_{\mu,\nu}-A_{\nu,\mu}$.

As mentioned earlier, the phase transition of the very early universe can be described by introducing a scalar Higgs field ϕ into the theory. One way to do this is to add an additional energy–momentum tensor $T'_{\mu\nu}$, due to the Higgs field, to the existing energy–momentum tensor on the right hand side of Einstein's equations (2.22). The form of this additional energy–momentum tensor is suggested by the Lagrangian of a scalar field, which is as follows (V is a suitable potential and we omit the factor g that makes this a scalar density – see section 2.2; this is taken in account in deriving (9.9 a,b)):

$$L=\tfrac{1}{2}\partial_{\mu}\phi\partial^{\mu}\phi-V(\phi). \tag{9.3}$$

It is well known (see, for example, Bogoliubov and Shirkov (1983, p. 17)) that for a scalar field the energy–momentum tensor associated with a Lagrangian L is given by

$$T_{\mu\nu}=\frac{\partial L}{\partial\phi_{,\mu}}\phi_{,\nu}-g_{\mu\nu}L. \tag{9.4}$$

For L given by (9.3) this gives

$$T'_{\mu\nu}=\partial_{\mu}\phi\partial_{\nu}\phi-g_{\mu\nu}L=\partial_{\mu}\phi\partial_{\nu}\phi-g_{\mu\nu}[\tfrac{1}{2}\partial_{\sigma}\phi\partial^{\sigma}\phi-V(\phi)]. \tag{9.5}$$

The energy–momentum tensor for a perfect fluid given by (2.23) for the comoving cosmological fluid can be written in the following form:

$$T^{\nu}_{\mu} = \mathrm{diag}(\varepsilon, -p, -p, -p), \tag{9.6}$$

where the tensor is written in matrix form with diagonal elements, other elements being zero. We now assume that the scalar field, ϕ, depends on the time t only, and if we write the tensor $T'_{\mu\nu}$ in the form (9.6), that is,

$$T'^{\nu}_{\mu} = \mathrm{diag}(\varepsilon', -p', -p', -p'), \tag{9.7}$$

we find from (9.5) the following relations for ε', p':

$$\varepsilon' = \tfrac{1}{2}\dot{\phi}^2 + V(\phi);\ p' = \tfrac{1}{2}\dot{\phi}^2 - V(\phi);\ \dot{\phi} \equiv \partial\phi/\partial t. \tag{9.8}$$

Thus the modified Einstein equations in the cosmological situation with the Higgs field ϕ are obtained by simply replacing ε by $\varepsilon + \varepsilon'$, and p by $p + p'$, with ε', p' given by (9.8).

There are some basic assumptions in this analysis which we must clarify. Firstly, we are dealing essentially with a region which is within the horizon distance at the time under consideration. This region, according to the above scenario, undergoes rapid expansion, cooling, etc., more or less independent of the rest of the universe. Yet we are using for this region the Robertson–Walker metric which is derived under the assumption that the entire space is homogeneous and isotropic. We are thus using the assumption here that the total space-time behaves in such a manner that the Robertson–Walker form of the metric is justified locally. Secondly, we are ignoring the spatial variation of ϕ, so that throughout the region ϕ takes a uniform value. This assumption leads to the relatively simple Equations (9.8). A third assumption, which is not a serious restriction, is that in (9.8) we have used the $k = 0$ form of the Robertson–Walker metric, that is, the form which has flat spatial geometry. We will do this throughout this chapter. Thus the Einstein equations are now given by (with $c = 1$):

$$(\dot{R}/R)^2 \equiv H^2 = (8\pi G/3)(\varepsilon + \varepsilon'), \tag{9.9a}$$

$$2\ddot{R}/R + H^2 = -8\pi G(p + p'), \tag{9.9b}$$

where ε', p' are given by (9.8) in terms of ϕ.

Consider now the situation in the very early universe when the temperature is higher than 10^{27} K. As mentioned earlier $\phi = 0$ so that from (9.8) we see that ε' has the constant value $V(0)$. On the other hand, if we assume that the equation of state is that of radiation, we see that R behaves like $t^{1/2}$ (see (4.47)) while ε behaves like t^{-2} (see (4.40)). Thus in (9.9a) the ε dominates the right hand side, so the evolution of R is as if the ϕ term did not

exist, that is, ε decreases like t^{-2} as t increases. Since the evolution of R is faster than the phase transition (for some set of parameters of Grand Unified Theories), ϕ remains at the value zero while the temperature goes below the critical. The ε term on the right hand side of (9.9a) becomes much less than ε', that is, $V(0)$, so that the evolution of R is given by

$$(\dot{R}/R)^2 = (8\pi G/3)V(0), \tag{9.10}$$

which has the solution

$$R = \exp(\zeta t), \quad \zeta^2 = (8\pi G/3)V(0), \tag{9.11}$$

provided $V(0)$ is positive, which is the case, for example, in Figs. 9.2 and 9.3. Thus the scale factor R undergoes exponential expansion. This behaviour is, in fact, that of de Sitter space, for which we give a little digression.

In (6.2a) and (6.2b), if we set $\varepsilon = p = k = 0$, we get (with $c = 1$):

$$(\dot{R}/R)^2 = \tfrac{1}{3}\Lambda, \tag{9.12a}$$

$$2\ddot{R}/R + (\dot{R}/R)^2 = \Lambda. \tag{9.12b}$$

It is readily verified that (9.12a) and (9.12b) are satisfied by

$$R = \exp(\tfrac{1}{3}\Lambda)^{1/2}t, \tag{9.13}$$

assuming that the cosmological constant is positive. The model given by (9.13) (with $k = 0$) is called the *de Sitter universe*, which is empty, has a positive cosmological constant, and has a non-trivial scale factor given by (9.13). Sometimes it is said that the de Sitter space represents 'motion without matter' as opposed to the Einstein universe (see (6.3)), which represents 'matter without motion'. Equation (9.13) gives the same behaviour as (9.11) and is also the form of the steady state universe, as mentioned earlier, which was put forward originally by Bondi and Gold (1948) and by Hoyle (1948). The latter is maintained at a steady state by the continuous creation of matter, the amount of which is cosmologically significant but negligible by terrestrial standards, so that no experiment on the conservation of mass is violated. Observations of the background radiation and others, however, contradict the steady state theory. It is curious that in the inflationary models one has to consider again a similar exponential metric (9.11), albeit for a very short period in the history of the universe.

When the energy–momentum tensor $T_{\mu\nu}$ of the cosmological fluid can be neglected in comparison with the energy–momentum tensor $T'_{\mu\nu}$ of the scalar field soon after the onset of the phase transition and when ϕ is still zero (that is, the situation that leads to the metric (9.11)), we see from (6.1)

(with $T_{\mu\nu}=0$) and (9.5), we get precisely the Einstein equations with the cosmological constant but zero pressure and density with $\Lambda=8\pi GV(0)$. Thus the cosmological constant reappears here in quite a different context.

Consider again the situation when the scalar field dominates but it has started deviating from zero. Using (9.8) and (9.9a) we get

$$(\dot{R}/R)^2\equiv H^2=(8\pi G/3)[\tfrac{1}{2}\dot{\phi}^2+V(\phi)]. \tag{9.14}$$

Consider the vanishing of the divergence of the energy–momentum tensor, which gives (3.79) for the Friedmann models, which we write here for convenience:

$$\dot{\varepsilon}+3(p+\varepsilon)\dot{R}/R=0.$$

In the present situation of the scalar field we should replace ε, p with ε', p' in this equation in accordance with (9.6), (9.7) and (9.8). Doing this, from (9.8) we get the following equation, after cancelling a factor $\dot{\phi}$:

$$\ddot{\phi}+2H\phi+V'=0,\ V'\equiv \mathrm{d}V/\partial\phi. \tag{9.15}$$

Equations (9.14) and (9.15) represent the equations which govern the evolution of the scale factor and the scalar field when the latter is the dominant agent of the evolution. In general, exact solutions are difficult to get for any reasonable form of the potential $V(\phi)$. For example, the forms depicted in Figs. 9.2 and 9.3 are given respectively by (9.16a) and (9.16b) below.

$$V(\phi)=\lambda_0\phi^2+\lambda_1\phi^3+\lambda_2\phi^4+V_0, \tag{9.16a}$$

$$V(\phi)=\lambda(\phi^2-\sigma^2)^2, \tag{9.16b}$$

for suitable values of the constants λ_0, λ_1, λ_2, V_0, λ, σ. However, it is very difficult to find exact solutions of (9.14), (9.15) for the forms (9.16a) and (9.16b) of the potential $V(\phi)$. In sections 9.4 and 9.6 we shall consider an exact solution found by the author (Islam, 2001a) for a potential $V(\phi)$ of the sixth degree. For V given by (9.16a) and (9.16b) one usually resorts to an approximation, in one form of which one ignores the $\ddot{\phi}$ term in (9.15), and takes $V(\phi)$ to be given by (9.16b), so that the system is initially at $\phi=0$ and 'rolls' slowly away from $\phi=0$, the speed of departure from $\phi=0$ gradually increasing (Brandenberger, 1987). However, these approximation schemes are unsatisfactory as they sometimes give ambiguous results. For example, Mazenko, Unruh and Wald (1985) argue that in many possible models, conditions for inflation do not obtain in the very early universe. This point of view has been opposed by Albrecht and Brandenberger

(1985) who also claim that there are many possible models in which a period of inflation does occur. It would probably be true to say that a completely satisfactory picture for inflation has not yet emerged. See also Pacher, Stein-Schabes and Turner (1987) and Page (1987). This was the situation about a decade ago in the late eighties; it has not changed substantially.

9.4 An exact inflationary solution

In this section we present an exact solution of the coupled scalar field cosmological equations (9.14) and (9.15), for $V(\phi)$ given as follows:

$$V(\phi) = V_0 + V_1\phi^2 + V_5\phi^5 + V_6\phi^6, \tag{9.17}$$

where the V_i are constants. The solution presented here does not, in fact, satisfy the properties appropriate to the inflation scenario that we have been discussing; for example, $\phi(0) \neq 0$ in this case. Nevertheless, it is of interest because it is an exact solution for a polynomial potential, probably the only exact solution known for such a potential (an exact solution for an exponential potential was found by Barrow, 1987), and the corresponding scale factor does have exponential behaviour over certain ranges of values of t. We will simply state the solution and verify that it is indeed a solution. In Section 9.6 we will generalize this solution.

Write $q = 3\pi G$, and let n be a constant. Choose the V_i as follows:

$$V_0 = 9n^2/8q;\ V_1 = n^2;\ V_5 = -\tfrac{2}{3}n^2q^{3/2};\ V_6 = 2n^2q^2/9. \tag{9.18}$$

The solution for ϕ is as follows:

$$\phi(t) = q^{-1/2}\exp(nt)[\exp(nt) + \xi]^{-1} \equiv q^{-1/2}x, \tag{9.19}$$

where ξ is a constant. It is readily verified that

$$\dot{\phi} = nq^{-1/2}(x - x^2);\ \ddot{\phi} = n^2q^{-1/2}(1 - 2x)(x - x^2). \tag{9.20}$$

With the use of (9.18) and (9.20), Equation (9.14) yields

$$\begin{aligned} H^2 &= \tfrac{8}{9}q(\tfrac{1}{2}\dot{\phi}^2 + V(\phi)) \\ &= n^2(1 + \tfrac{4}{3}x^2 - \tfrac{8}{9}x^3 + \tfrac{4}{9}x^4 - \tfrac{16}{27}x^5 + \tfrac{16}{81}x^6) \\ &= n^2(1 + \tfrac{2}{3}x^2 - \tfrac{4}{9}x^3)^2, \end{aligned} \tag{9.21}$$

so that

$$H = \pm n(1 + \tfrac{2}{3}x^2 - \tfrac{4}{9}x^3). \tag{9.22}$$

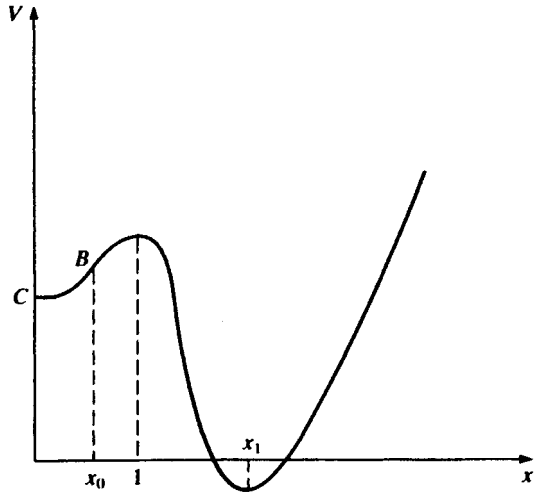


Fig. 9.5. This diagram depicts the form of the potential given by (9.17) and (9.18) in terms of x (9.19). The curve has turning points at $x=0$, $x=1$ and $x=x_1$, where x_1 is slightly greater than 2. At $t=0$, x starts at x_0 and goes to zero as t tends to infinity; $x_0=(1+\xi)^{-1}$. The potential is negative for a finite range of values of the field, but the negative portion does not come into play for the particular solution found here. One gets more realisitic behaviour for the generalized solution (see Fig. 9.7).

If we now substitute for $\ddot{\phi}$, $\dot{\phi}$, H and V' from (9.17), (9.18), (9.20) and (9.22) into (9.15) we find, after some reduction, that the latter is satisfied. The scale factor $R(t)$ can be determined by integrating (9.22), where the negative sign must be taken to satisfy (9.15). The result of the integration is as follows:

$$R(t)=A\exp(-nt)[\exp(nt)+\xi]^{-2/9}\exp\{(2\xi/9)\exp(nt)[\exp(nt)+\xi]^{-2}\}, \tag{9.23}$$

where A is an arbitrary constant. If we take n to be negative, R has an exponential increase, while $\phi(t)$ goes to zero as t tends to infinity.

The form of the potential (9.17) is that depicted in Fig. 9.5. This can be seen from the fact that the equation $V'(\phi)=0$ determining the turning points is given as follows:

$$\phi(2V_1+5V_5\phi^3+6V_6\phi^4)=0. \tag{9.24}$$

With the use of (9.18), in addition to the root at $\phi=0$ we get the following equation (in terms of $x=q^{1/2}\phi$):

$$2x^4-5x^3+3=0, \tag{9.25}$$

which has only two real roots, one at $x = 1$, and the other at slightly above $x = 2$. If we assume ξ to be positive, at $t = 0$, x has the value $(1 + \xi)^{-1}$, and it tends towards zero (for negative n) as t tends to infinity. Thus it goes 'down' the slope from the point B to the point C in Fig. 9.5.

Before closing this section, we will state briefly Barrow's (1987) solution. This is of the form

$$R(t) = R_0(t/t_0)^b, \tag{9.26a}$$

$$\phi(t) = \phi_0(\log t/\log t_0), \tag{9.26b}$$

$$V(\phi) = V_0 \exp(-\lambda\phi), \tag{9.26c}$$

where t_0, R_0, ϕ_0, V_0, b and λ are suitable constants. Note that this solution gives a power law inflation. In Section 9.6 we will consider more inflationary solutions, including one corresponding to a generalized form of the simpler solution which will make it easier to follow the generalization.

9.5 Further remarks on inflation

In Sections 9.3 and 9.4 we have attempted to give a quantitative description of inflation. It is indicated how at the onset of the phase transition a de Sitter-like exponential expansion might occur due to the presence of the Higgs field, represented here by a scalar field. However, the further evolution of the scale factor $R(t)$ and the scalar field are not clear due to the difficulty of obtaining exact solutions of the coupled equations, (9.14) and (9.15), for any reasonable potential. In the last section we obtained an exact solution which, though somewhat unrealistic, offers some hope that physically more meaningful solutions might be found. Its generalization in Section 9.6 is more realistic.

To discuss phase transitions in the very early universe one must know the so-called 'effective potential' $V(\phi, T)$ as a function of the scalar field ϕ and the temperature T. For temperatures above the critical temperature for the phase transition, the symmetric phase ($\phi = 0$), in which the symmetry among the various interactions is manifest, is the global minimum of the effective potential. One has to derive the effective potential from quantum field theoretic considerations (see, for example, Brandenberger (1985)), but even here one has to resort to approximation schemes. The effective potential used by Albrecht and Steinhardt (1982) and by Linde (1982) is based on the Coleman–Weinberg mechanism (Coleman and Weinberg, 1973) and is given as follows:

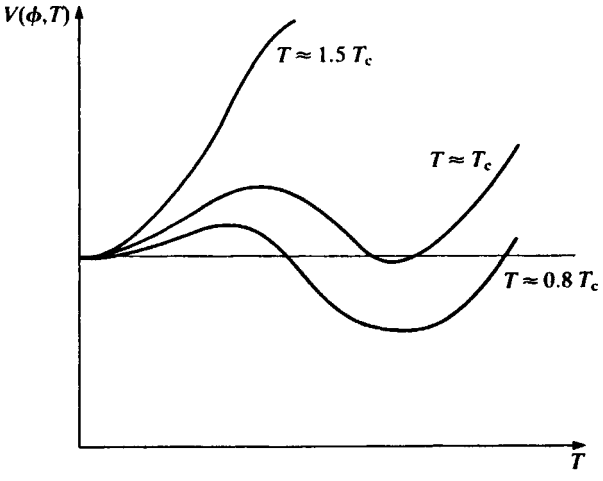


Fig. 9.6. This figure gives the potential represented by (9.27) for three values of the temperature, one slightly above the critical temperature for the phase transition T_c, one near T_c, and a third slightly below T_c. (This is a simplified form of the diagram given in Albrecht and Steinhardt (1982).)

$$V(\phi, T) = \alpha\phi^2 - A\phi^4 + B\phi^4 \log(\phi^2/\phi_0^2) + 18(T^4/\pi^2) \times \int_0^\infty \mathrm{d}y y^2 \log\{1 - \exp[-(y^2 + 25g^2\phi^2/8T^2)^{1/2}]\}, \qquad (9.27)$$

where α, A, B, ϕ_0, g are suitable constants. Figure 9.6 depicts the effective potential given by (9.27) for three typical values of the temperature T.

We will now give some tentative answers to the questions raised at the end of Section 9.2. As regards the monopole problem, the new inflationary model attempts to solve the horizon, magnetic monopole and domain-wall problems in one stroke, namely, by the requirement that before the phase transition the region or space from which the observable universe evolved was much smaller than the horizon distance, so that this region had time to homogenize itself, and because of the inflation from a small portion, the observable universe is expected to have very few monopoles and domain walls, consistent with observation. As regards the matter–antimatter asymmetry, it is possible in some forms of Grand Unified Theories to produce the observed excess of matter over antimatter by elementary particle interactions at temperatures just below the critical temperature of the phase transition, provided certain parameters are suitably chosen.

However, there are still many uncertainties in this analysis, but the very possibility of deriving the asymmetry is interesting.

In the form in which the model of the inflationary universe was originally proposed by Guth in 1981 it had a serious defect in that the phase transition itself would have created inhomogeneities to an extent which would be inconsistent with those observed at present. The difficulties with the new inflationary models are, firstly, as already indicated, a completely satisfactory quantitative treatment does not as yet exist and, secondly, in the approximate treatment of the slow-rollover transition, one requires fine tuning of the parameters which seems somewhat implausible. A great deal of further work needs to be done to clarify the above questions.

We will mention briefly some related points of some importance which we have not been able to deal with in detail. These are firstly the origin of inhomogeneities in the universe that we observe today. One aspect of this problem is similar to the horizon problem. One way to this problem is to consider the inhomogeneities at any time as consisting of perturbations to the smooth background which involve wavelengths of all scales. However, as one extrapolates this analysis to earlier times, a certain range of the larger wavelengths becomes longer than the horizon distance, and it becomes a problem as to how these larger wavelengths arose; in other words, one has to find a mechanism in which wavelengths larger than the horizon distance are excluded at any time. Another related aspect is the formation of one-dimensional defects during the phase transition; these are known as *cosmic strings* and may have a role to play in the formation or origin of galaxies. We refer the reader to Brandenberger (1987) and Rees (1987) for these questions. Press and Spergel (1989) in particular, explain in a picturesque manner how a field-theoretic description of matter (such as that given by the Lagrangian (9.3)) implies fossilized one-dimensional remnants of an earlier, high-temperature phase. Different symmetries of the Langrangians describing possible states of matter in the very early universe give rise to different kinds of remnants, which arise from certain invariant topological properties of space-time. For other consequences of the phase transition in the very early universe, the reader is referred to Miller and Pantano (1989) and to Hodges (1989); the latter author considers domain wall formation. The question of chaotic inflation is considered by Futamase and Maeda (1989) and by Futamase, Rothman and Matzner (1989). Adams, Freese and Widrow (1990) study the problem of the evolution of non-spherical bubbles in the very early universe. The problem of the formation of clusters of galaxies from cosmic strings is investigated by Shellard and Brandenberger (1988). Lastly, we mention

'extended inflation' (La and Steinhardt, 1989; La, Steinhardt and Bertschinger, 1989) in which a special phase transition is not needed, that is, $V(\phi)$ can have a significant barrier between the true and false vacuum phases. Steinhardt (1990) shows that this model accommodates initial conditions leading to $\Omega \leqslant 0.5$. In 'extended inflation' the defects of 'old inflation' are avoided if the effective gravitational constant, G, varies with time during inflation.

9.6 More inflationary solutions

Ellis and Madsen (1991) find a number of exact cosmological solutions with a scalar field and non-interacting radiation, which could provide some new inflationary models. They give a method of 'generating' a class of solutions, following an old idea due to Synge (1955). We give two examples, in which the radiation density is set equal to zero, so that one has a pure scalar field (with $k = 1$, i.e. flat spatial sections):

$$R(t) = A \exp(wt),\ A,\ w \text{ constant} > 0, \tag{9.28a}$$

$$\phi(t) = \phi_0 \pm (B/w)\mathrm{e}^{-wt},\ \phi_0 = \text{constant},\ B^2 = (4A^2\pi G)^{-1}, \tag{9.28b}$$

$$V(\phi) = (3w^2/8\pi h) + w^2(\phi - \phi_0)^2. \tag{9.28c}$$

This solution gives the usual de Sitter exponential expansion, without a singularity in the finite past, unlike the following solution, which has a singularity in the finite past:

$$R(t) = A \sinh(wt),\ A,\ w \text{ constant} > 0, \tag{9.29a}$$

$$\phi(t) = \phi_0 \pm (B/w) \log\left(\frac{\mathrm{e}^{wt} - 1}{\mathrm{e}^{wt} + 1}\right),\ \phi_0 = \text{constant},$$

$$B^2 = \frac{1}{4\pi G}\left(w^2 + \frac{k}{A^2}\right) \geqslant 0, \tag{9.29b}$$

$$V(\phi) = \frac{3w^2}{8\pi G} + B^2\left\{\sinh\left[\frac{2w}{\mathrm{B}}(\phi - \phi_0)\right]\right\}^2. \tag{9.29c}$$

In (9.29b) the inequality is true for $k = 0$, 1 and can always be satisfied for $k = -1$. Ellis and Madsen discuss various properties of these solutions in the context of inflationary models. An interesting aspect of these solutions is that they allow a wide variety of behaviour for the density parameter Ω.

A number of interesting power law and exponential inflationary

solutions, including ones which have intermediate expansion rates, have been considered by Barrow (1990), Barrow and Maeda (1990) and Barrow and Saich (1990).

We will now consider a generalization of the exact solution found in Section 9.4. Some of the unphysical features of the previous solution are rectified in the new solution. We first describe the solution and discuss its properties. We verify in the Appendix to this chapter that it is indeed a solution, and derive some of the properties. The potential (9.17) is generalized as follows:

$$V(\phi) = V_0 + V_1\phi + V_2\phi^2 + V_3\phi^3\ V_4\phi^4 + V_5\phi^5 + V_6\phi^6, \tag{9.30}$$

where the V_i are constants, which are expressed in terms of three constants α, β and n as follows:

$$\begin{aligned}
&V_0 = q^{-1}\{\alpha^2 - \tfrac{1}{2}n^2\beta^2(1-\beta)^2\};\ V_1 = q^{-\frac{1}{2}}\beta(\beta-1)\{2\sqrt{2}n\alpha - n^2(2\beta-1)\};\\
&V_2 = \{\sqrt{2}n(2\beta-1)\alpha + n^2(-\tfrac{1}{2} + 3\beta - \beta^2 - 4\beta^3 + 2\beta^4)\};\\
&V_3 = q^{\frac{1}{2}}\{(2\sqrt{2}/3)n\alpha + n^2(1 - 6\beta^2 + 4\beta^3);\ V_4 = (10/3)qn^2\beta(\beta-1);\\
&V_5 = (2/3)q^{3/2}n^2(2\beta-1);\ V_6 = (2/9)q^2n^2.
\end{aligned} \tag{9.31}$$

The function $\phi(t)$ is given by the following relation:

$$\phi(t) = q^{-\frac{1}{2}}\{-\beta + e^{nt}(e^{nt} + \xi)^{-1}\}, \tag{9.32}$$

where ξ, as before, is also a constant. This solution reduces to the previous one if $\beta = 0$ and $\alpha = -(3/2\sqrt{2})n$. The corresponding relation for $R(t)$ will be given in the Appendix.

We discuss some properties of the new solution. First note that $\phi(t)$ need not be non-zero at $t = 0$. In fact $\phi(0) = 0$ if β is chosen to be equal to $(1+\xi)^{-1}$. Besides, instead of the unusual behaviour of the potential depicted in Fig. 9.5, one gets a variety of more realistic possibilities, which are somewhat like the behaviour of the Coleman–Weinberg potential displayed in Fig. 9.6, for different values of the constant β, which thus behaves like the temperature T. To see this, we set $V_1 = 0$, so that $dV/d\phi$ vanishes at $\phi = 0$. This is achieved by taking α to be given as follows, in terms of β and n: $\alpha = n(2\beta-1)/2\sqrt{2}$. Henceforth we use this value of α. For the purpose of drawing the potential curves for different values of β, we define a modified potential U, proportional to V, and express it as a polynomial in y, defined as follows:

$$U = qV/n^2;\ y = q^{\frac{1}{2}}\phi = x - \beta, \tag{9.33}$$

where x is defined by (9.19), that is, $x = e^{nt}(e^{nt} + \xi)^{-1}$. After some reduction we find the following expression for U in terms of y:

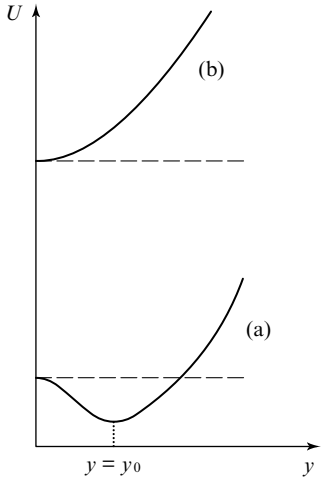


Fig. 9.7. The potential $U(y; \beta)$ for two values of β, given respectively by (9.40) and (9.43) ((a) and (b) respectively). Curve (a) is similar to that in Fig. 9.3 and curve (b) is similar to one of the curves in Fig. 9.6.

$$U = \{(-4\beta^4 + 8\beta^3 - 4\beta + 1)/8 + \beta(\beta - 1)\,(2\beta^2 - 2\beta - 1)y^2 + [4\beta^3 - 6\beta^2 + (2/3)\beta - 1/3]y^3 + (10/3)\beta(\beta - 1)y^4 + (2/3)\,(2\beta - 1)y^5 + (2/9)y^6\}. \tag{9.34}$$

The potential U can thus be written as $U(y, \beta)$, that has different functional form for various values of β. The qualitative behaviour of U for two different values of β is displayed in Fig. 9.7, which is not drawn to scale. If we ignore the different starting values of U (for $y = 0$), we see that the curves are somewhat like two of those in Fig. 9.3 and Fig. 9.6. Some details of the derivation are given in the Appendix. The new solution is thus more realistic, in that there are more parameters which give a variety of behaviours for the potential. A further generalization to an eighth degree potential has been considered by the author in collaboration (Azad and Islam, 2001). Although these polynomial potentials differ in form from more realistic potentials such as (9.27), the higher the degree of the polynomial potential, the closer it can be made to any desired function. In this sense potentials of higher order polynomials are useful in this context, as approximations to the Coleman–Weinberg or other potentials (see, e.g., Lazarides, 1997). The corresponding $R(t)$ is not so easy to ascertain. Some unphysical features remain in the new solution, such as the fact that, if one insists on $\phi(0) = 0$, as t tends to infinity, both for positive and negative n, $\phi(t)$ tends to a (negative) constant. Nevertheless, the fact that

one can derive an exact and explicit solution for a fairly complicated potential enables one to examine some aspects in detail. The new solution was obtained by the author some years ago when starting to prepare this second edition. It is presented here for the first time. It is hoped to consider some incomplete aspects and related matters in a future work. In this connection mention may be made of the work of Barrow and Liddle (1997), Barrow (1993), Rahaman and Rashid (1996), and Rahaman (1996).

Appendix to Chapter 9

In this Appendix we present the main steps of the derivation of the new solution and some of its properties. The method is essentially the same as in Section 9.4 with the restricted potential (9.17), but the steps are more elaborate. The basic equations, as before, are (9.14) and (9.15). The equations (9.20) remain unchanged, except that it is more convenient to express these, and other relations involving x, in terms of $y = x - \beta$. Following steps similar to those given in (9.21) and (9.22), one finds that the form (9.31) of the potential leads to an expression for H^2 (given by (9.14) or the top equation in (9.21)) in terms of y that is a perfect square, which leads to the following relation for H:

$$\begin{aligned} H &= \{(4/9)ny^3 + (2/3)n(2\beta-1)y^2 + (4/3)n\beta(\beta-1)y + (2\sqrt{2}/3)\alpha\} \\ &= \{(4/9)nq^{3/2}\phi^3 + (2/3)n(2\beta-1)q\phi^2 + (4/3)\beta(\beta-1)nq^{\frac{1}{2}}\phi + (2\sqrt{2}/3)\alpha\}. \end{aligned} \tag{9.35}$$

To verify (9.15), we use the relations for $\dot{\phi}$ and $\ddot{\phi}$ given by (9.20), expressed in terms of y, and insert the expressions for the V_i given by (9.31) in the derivative:

$$\mathrm{d}V/\mathrm{d}\phi = V_1 + 2V_2\phi + 3V_3\phi^2 + 4V_4\phi^3 + 5V_5\phi^4 + 6V_6\phi^5. \tag{9.36}$$

The resulting expression for the left hand side of (9.15) is as follows:

$$\begin{aligned} &\ddot{\phi} + 3H\dot{\phi} + \mathrm{d}V/\mathrm{d}\phi = nq^{-\frac{1}{2}}\{n(y+\beta)\,(1-2y-2\beta)\,(1-y-\beta) \\ &+ 3(y+\beta)\,(1-y-\beta)\,[(4/9)ny^3 + (2/3)n(2\beta-1)y^2 \\ &+ (4/3)n\beta(\beta-1)y + (2\sqrt{2}/3)\alpha] + \beta(1-\beta)\,[-2\sqrt{2}\alpha + n(2\beta-1)] \\ &+ 2[\sqrt{2}\alpha(2\beta-1) + n(-\tfrac{1}{2} + 3\beta - \beta^2 - 4\beta^3 + 2\beta^4)]y \\ &+ 3[(2\sqrt{2}/3)\alpha + n(1 - 6\beta^2 + 4\beta^3)]y^2 + (40/3)n\beta(\beta-1)y^3 \\ &+ (10/3)n(2\beta-1)y^4 + (4/3)ny^5\}. \end{aligned} \tag{9.37}$$

It can be verified that the right hand side of (9.37) vanishes identically. The fact that the term independent of y and the coefficient of y^5 vanish can be

seen by inspection. Thus both the equations (9.14) and (9.15) are satisfied. To obtain $R(t)$ it is easier to express H in terms of $x(=e^{nt}(e^{nt}+\xi)^{-1})$, as follows:

$$H=\{n(4/9)x^3-n(2/3)x^2+(2\sqrt{2}/3)\alpha+n(2/3)\beta^2-n(4/9)\beta^3\}. \quad (9.38)$$

Compare this with (9.22), where the lower sign is the relevant one. When $\beta=0$ and $\alpha=-(3/2\sqrt{2})n$, (9.38) reduces to (9.22). Noting that $H=\dot{R}/R=(\mathrm{d}/\mathrm{d}t)(\log R)$, one can integrate (9.38) as in (9.22), to get the following expression for $R(t)$ (with B an arbitrary constant):

$$R(t)=B\exp(kt)\,[\exp(nt)+\xi]^{-2/9}\exp\{(2\xi/9)\exp(nt)\,[\exp(nt)+\xi]^{-2}\}, \quad (9.39)$$

where $k=(2\sqrt{2}/3)\alpha+(2/9)n\beta^2(3-2\beta)$, so that $k=-n$, as in (9.23) when $\beta=0$ and $\alpha=-(3/2\sqrt{2})n$. Thus the only difference between the expressions (9.23) and (9.39) for $R(t)$ is that in the latter the first factor is $\exp(kt)$ instead of $\exp(-nt)$, with k given as above. One can choose values of α and β so that k is positive and, for a period at least, there is exponential expansion.

We now consider the form of the potential function $U(y,\beta)$ given by (9.34) for two specific values of β, namely, $\beta=(\sqrt{3}+1)/2\approx 1.366$, and $\beta=3/2$. We choose these values because they lead to interesting behaviour for the potential, and it is possible to determine this behaviour analytically without resorting to numerical computation.

Consider $\beta=(\sqrt{3}+1)/2$ first. For this value of β the function U is given as follows:

$$U(y;(\sqrt{3}+1)/2)=\{1/8-((\sqrt{3}-1)/\sqrt{3})y^3+(5/3)y^4+(2/\sqrt{3})y^5+(2/9)y^6\}. \quad (9.40)$$

Note that the coefficient of y^2 vanishes and that of y^3 is negative. The turning points of this curve, apart from the one at $y=0$, occur at the roots of the following cubic, obtained by setting $\mathrm{d}U/\mathrm{d}y=0$ and cancelling a factor y^2:

$$(4/3)y^3+(10/\sqrt{3})y^2+(20/3)y-3+\sqrt{3}=0. \quad (9.41)$$

It can be verified that the left hand side of (9.41) is negative for $y=0.2$ and positive for $y=0.25$. There is thus a root of this equation at $y=y_0$ with $0.2<y_0<0.25$. Furthermore, the left hand side of (9.41) is clearly an increasing function of y for positive y. The root $y=y_0$ is therefore the only positive root. We now show that the function U given by (9.40) is positive for positive y. There are values of β for which some parts of the potential

(9.34) are negative, as is clear from the earlier restricted potential depicted in Fig. 9.5. The negative parts may or may not be unphysical; to avoid such questions we choose β for which the potential does not have negative parts; this also makes a better analogy with potentials shown in Fig. 9.6. Now U given by (9.40) takes its minimum value for $y = y_0$, so that if it is positive at this point, it will be positive for all $y > 0$. Consider the value $y = y_1 = \sqrt{3}(\sqrt{3} - 1)/5$, for which the cubic and quartic terms in (9.40) cancel, and U is clearly positive at this value: $y_1 \approx 0.254$. It can also be verified that the first two terms in (9.40) are positive for $y = y_1$:

$$1/8 - [(\sqrt{3} - 1)/\sqrt{3}]y_1^3 > 0. \tag{9.42a}$$

Now a necessary, but not sufficient condition for U to be negative at $y = y_0$ is that the first two terms must be negative (because the other terms are > 0):

$$1/8 - [(\sqrt{3} - 1)/\sqrt{3}]y_0^3 < 0. \tag{9.42b}$$

Now these first two terms form a decreasing function of y (for $y > 0$), and $y_1 > y_0$, so we have a contradiction in (9.42b). Hence (9.42b) cannot be right and so U must be positive for $y = y_0$, and so positive for all $y > 0$.

Consider next $\beta = 3/2$. From (9.34) we get in this case:

$$U(y; 3/2) = \{11/16 + (3/8)y^2 + (2/3)y^3 + (5/2)y^4 + (4/3)y^5 + (2/9)y^6\} \tag{9.43}$$

This curve has no turning points for $y > 0$, since $\mathrm{d}U/\mathrm{d}y$, having all positive terms, cannot vanish for $y > 0$. The behaviour of $U(y; \beta)$ for the two values of β are thus as depicted in Fig. 9.7. From the analysis carried out here it is also clear that a value of β is likely to exist for which $U(y; \beta)$ has the form displayed in Fig. 9.3, that is, the 'slow roll over' form; (9.40) has some similarity to this.

10
Quantum cosmology

10.1 Introduction

We saw in the previous chapters that the standard model predicts a singularity sometime in the past history of the universe where the density tends to infinity. In Chapter 7 we also saw there is reason to believe that the existence of singularities may not be a feature peculiar to the highly symmetric Friedmann models, but may exist in any general solution of Einstein's equations representing a cosmological situation. Many physicists think that the existence of singularities in general relativity is unphysical and points to the breakdown of the theory in the extreme situations that singularities purport to represent. Indeed, in these extreme conditions the quantum nature of space-time may come into play, and there have been suggestions that when the quantum theory of gravitation is taken into account, singularities may not arise. However, the quantization of gravitation is notoriously difficult – there does not, at present, exist any satisfactory quantum theory of gravitation, whether the gravitation theory is general relativity or any other reasonable theory of gravity. However, there have been some approximate schemes to try and answer at least partially some of the questions that a quantum theory of gravitation is supposed to answer. One of these schemes is quantum cosmology. We shall only give a brief and incomplete account of quantum cosmology in this chapter, as the technicalities are mostly beyond the scope of this book. This chapter is based mainly on Hartle and Hawking (1983), Hartle (1984, 1986), Narlikar and Padmanabhan (1983), and Islam (1993, 1994).

We give first a very simple-minded description of quantum theory and see what kind of light an extension of this theory to the cosmological situation may be expected to throw. The quantum theory is supposed to be the basic and fundamental theory which describes all physical phenomena. Classical

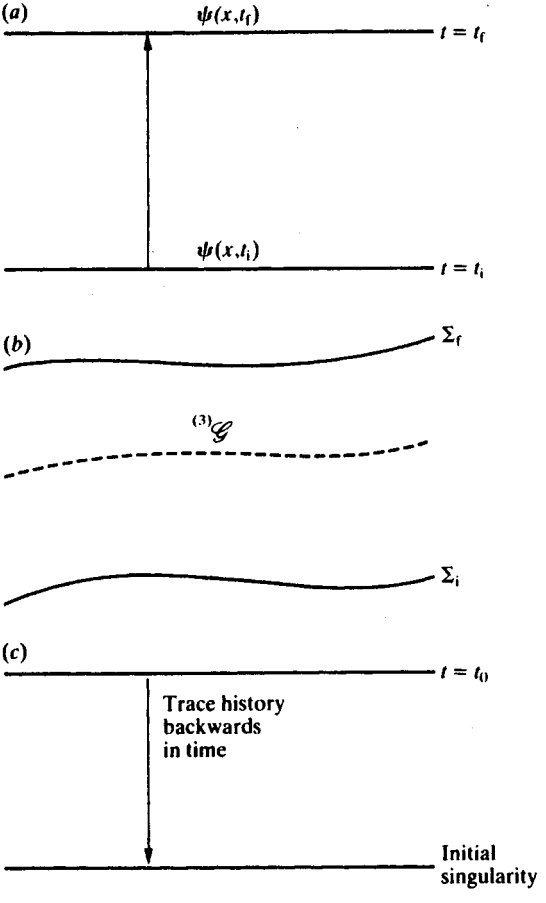


Fig. 10.1. (*a*) In the typical quantum mechanical situation the state of the system is given by the wave function ($\psi(x, t)$) at time $t = t_i$; the Schrödinger equation gives the wave function at a later time $t = t_f$. (*b*) In general relativity the three-geometry is given on a space-like hypersurface Σ_i; the quantum theory of gravitation then gives the probability for the three-geometry on the hypersurface Σ_f. (*c*) If the state of the universe is known at the present time t_0, the theory should predict the probability of different states that are possible near the 'singularity'.

(non-quantum) physics, including general relativity, is supposed to be an approximation in situations where the action becomes large compared to Planck's constant h, which is of the order of 10^{-27} erg s. The description we will give here is somewhat crude, but it has the merit of putting in a nutshell the kind of approach the quantum cosmologist has in mind. Consider Fig. 10.1, where (*a*) represents a simple quantum mechanical system, which is

described by a single spatial coordinate x at any time t. The wave function $\psi(x, t)$ represents the quantum mechanical state of the system at time t. We will not define here completely what we mean by the 'quantum mechanical state', but suffice it to say that if the wave function is known, all questions of physical interest can be answered with the use of the wave function. It is well known that if the wave function is known at a certain time t_i, the Schrödinger equation then enables us to calculate the wave function at a later time t_f. Extending the simple description given here to the case of a space-time geometry, we might suppose that space-time evolves from a space-like hypersurface Σ_i to another space-like hypersurface Σ_f; the condition of the space-time on any space-like hypersurface is given by the three-geometry on the hypersurface. For example, if we choose coordinates such that the metric may be written as follows

$$ds^2 = dt^2 + g_{jk}\, dx^j\, dx^k, \quad j, k = 1, 2, 3, \tag{10.1}$$

so that the space-like hypersurfaces may be taken as $t =$ constant, the three-geometry ${}^{(3)}\mathcal{G}$ at any time is given by the values of g_{jk} at that time. The evolution of the space-time from Σ_i to Σ_f is given classically by Einstein's equations, but in a quantum mechanical description we may ask for the probability of the space-time having any given three-geometry at Σ_f if it has a certain given three-geometry on Σ_i. In practice many difficulties arise; for example, there is the freedom on any surface $t =$ constant (in the metric (10.1)) of carrying out a purely spatial coordinate transformation which does not affect the intrinsic geometry, and one must disentangle the effects of such transformations from the physical evolution of the space-time. The manner in which such a probability amplitude might be found we will consider later. Finally (see (*c*) of Fig. 10.1), if we did find such a description, we could use the result backwards from the state of the universe at any time t_0 to a period of time near the supposed singularity, to study the quantum mechanical nature of the universe near such a point.

10.2 Hamiltonian formalism

As mentioned earlier, there exists as yet no satisfactory theory for quantizing gravitation. One of the approaches tried so far is the Hamiltonian approach. As is well known, in the Hamiltonian approach one has to single out a particular time coordinate. In general relativity this corresponds to choosing a particular manner of 'slicing' space-time with space-like hypersurfaces. We first give an elementary discussion of the non-relativistic Hamiltonian formalism and the corresponding derivation

of the Schrödinger equation: this is mainly to have a simple situation in mind while tackling the more complicated situation later.

We start with a Lagrangian $L(q, \dot{q})$ depending on a generalized coordinate q and its time derivative $\dot{q}$. The equation of motion is found by varying the action S derived from L given as follows:

$$S = \int_{t_1}^{t_2} L(q, \dot{q})\, \mathrm{d}t. \tag{10.2}$$

The condition that the variation $q(t) \to q(t) + \delta q(t)$, with $\delta q(t_1) = \delta q(t_2) = 0$ gives $\delta S = 0$ leads to the Euler–Lagrange equation of motion:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{q}}\right) - \frac{\partial L}{\partial q} = 0. \tag{10.3}$$

Corresponding to the coordinate q one defines the generalized momentum p as follows:

$$p = \partial L / \partial \dot{q}. \tag{10.4}$$

One then eliminates $\dot{q}$ in favour of p with the use of (10.4), and defines the Hamiltonian as follows:

$$H(p, q) = p\dot{q} - L(q, \dot{q}), \tag{10.5}$$

where it is assumed that $\dot{q}$ has been expressed in terms of p and q. From (10.3)–(10.5) it is readily seen that

$$\dot{p} = -\partial H / \partial q, \quad \dot{q} = \partial H / \partial p. \tag{10.6}$$

One defines the Poisson bracket of two functions F, G of p, q as follows:

$$\{F, G\} = \frac{\partial F}{\partial q}\frac{\partial G}{\partial p} - \frac{\partial F}{\partial p}\frac{\partial G}{\partial q}, \tag{10.7}$$

so that (10.6) can be written as follows:

$$\dot{p} = \{p, H\}, \quad \dot{q} = \{q, H\}. \tag{10.8}$$

It is well known that in quantum mechanics the variables q, p, H, etc., become operators in such a manner that the Poisson bracket can be replaced by commutators as follows ($\hbar = h/2\pi$):

$$\{q, H\} \to [q, H] = (\mathrm{i}\hbar)^{-1}(qH - Hq), \tag{10.9}$$

etc., so that, by comparing with the equation of motion of a free particle derived from the Lagrangian $L = \frac{1}{2}m\dot{q}^2$ (given by $\dot{q} = p/m$), we have the following commutation relation between q and p in quantum mechanics:

$$qp - pq = \mathrm{i}\hbar. \tag{10.10}$$

This implies that p can be expressed as the operator $p = -\mathrm{i}\hbar\,\partial/\partial q$. One can also show that the energy E can be replaced by the operator $\mathrm{i}\hbar\,\partial/\partial t$. It is readily seen that if the Lagrangian is given by

$$L = \tfrac{1}{2}m\dot{q}^2 - V(q), \tag{10.11}$$

which represents a particle moving in a potential $V(q)$, the Hamiltonian is given by

$$H = (2m)^{-1}p^2 + V(q), \tag{10.12}$$

so that, since the Hamiltonian represents the energy, we get the Schrödinger equation by applying both sides of (10.12) as operators to the wave function $\psi(q, t)$, whose modulus square $|\psi(q, t)|^2$ represents the probability density of finding the particle at q at time t; in fact $|\psi|^2\,\mathrm{d}q$ is the probability of the particle being between q and $q + \mathrm{d}q$:

$$\mathrm{i}\hbar\frac{\partial\psi}{\partial t} = \left[-\frac{h^2}{2m}\frac{\partial^2}{\partial q^2} + V(q) \right]\psi. \tag{10.13}$$

Consider now the Lagrangian for several particles given by

$$L = \sum_r L(q_r, \dot{q}_r), \tag{10.14}$$

where q_r represents the coordinate of the rth particle, and the generalized canonical momentum corresponding to q_r is given by

$$p_r = \partial L/\partial\dot{q}_r. \tag{10.15}$$

The corresponding Hamiltonian is given by

$$H(p_r, q_r) = \sum_r p_r\dot{q}_r - \sum_r L(q_r, \dot{q}_r). \tag{10.16}$$

We now consider the case of the Lagrangian of a field – this is like replacing the coordinate of the rth particle $q_r(t)$ by $\phi(\mathbf{x}, t)$, so that the index r is replaced by the spatial position $\mathbf{x}$ in a suitably limiting sense. The Lagrangian in this case is a function of the fields $\phi(\mathbf{x}, t)$ and the time and spatial derivatives $\partial_\mu\phi(\mathbf{x}, t)$. The sum over particles becomes an integral over the spatial coordinates ($\mathscr{L}$ is the Lagrangian density):

$$L(t) = \int \mathscr{L}(\phi, \partial_\mu\phi)\mathrm{d}^3x. \tag{10.17}$$

One defines a field $\pi(\mathbf{x}, t)$ canonically conjugate to $\phi(\mathbf{x}, t)$ as follows:

$$\pi = \partial\mathscr{L}/\partial\dot{\phi}. \tag{10.18}$$

In analogy with the particle case one defines the Hamiltonian:

$$H(t) = \int (\pi\dot{\phi} - \mathscr{L})\, \mathrm{d}^3x. \tag{10.19}$$

One can also make a simple-minded extension of the idea of a wave function to that of a wave functional $\Psi[\phi, t]$ which is a functional of the fields and a function of the time. It can be interpreted as saying that $|\psi(\phi, t)|^2\delta\phi$ is the probability of finding the field configuration between ϕ and $\phi+\delta\phi$ at time t. In analogy with the particle case, the Schrödinger equation in this case is given by

$$H\Psi[\phi, t] = \mathrm{i}\hbar\, \partial\Psi/\partial t. \tag{10.20}$$

Here H is given by (10.19), where the $\dot{\phi}$ has to be eliminated in favour of π using (10.18), and later π has to be replaced by the operator $-\mathrm{i}\hbar\, \delta/\delta\phi$, that is, $-\mathrm{i}\hbar$ times the functional derivative with respect to ϕ.

A functional is a number which depends on a function on the whole domain of its definition. Restricting to one variable x, a functional F of a function $A(x)$ may be given by

$$F[A] = \int f(x)A(x)\, \mathrm{d}x, \tag{10.21}$$

where $f(x)$ is a fixed function. The fundamental relation for taking functional derivatives is the following one:

$$\delta A(x)/\delta A(x') = \delta(x-x'), \tag{10.22}$$

where on the right we have the Dirac delta function. With the use of (10.22) we find readily that

$$\frac{\delta F}{\delta A(x')} = \int f(x)\, \frac{\delta A(x)}{\delta A(x')}\, \mathrm{d}x = \int f(x)\delta(x-x')\mathrm{d}x = f(x'). \tag{10.23}$$

In a similar manner with the use of (10.22) and the rules of ordinary differentiation, one can evaluate all functional derivatives.

The Wheeler–De Witt equation is a functional differential equation. The reader may not be familiar with functional derivatives and the corresponding functional differential equations. We have given above a very incomplete and somewhat crude account of the topic. In the next two or three sections we give a more detailed discussion. The subject matter may be unfamiliar and difficult, so the uninterested reader can skip these sections, but we believe those who are keen on quantum cosmology may find them of some interest. These are taken mainly from Islam (1994).

We first discuss (10.18), (10.19) and (10.20) for a specific Lagrangian and corresponding Hamiltonian.

10.3 The Schrödinger functional equation for a scalar field

We start with the following Lagrangian density for the scalar field ϕ, where m is the mass, $U(\phi)$ a suitable non-linear function of ϕ, for example, a polynomial of degree greater than 2, and the derivatives have their usual meaning:

$$\mathscr{L}=\tfrac{1}{2}\partial_\mu\phi\partial^\mu\phi-\tfrac{1}{2}m^2\phi^2-U(\phi),\quad \partial_\mu\phi=\partial\phi/\partial x^\mu, \text{ etc.} \tag{10.24}$$

The equation of motion, or field equation, is as follows:

$$(\Box+m^2)\phi+U'(\phi)=0,\quad \Box\equiv\partial_\mu\partial^\mu,\quad U'\equiv dU/d\phi. \tag{10.25}$$

Setting $c=1$, we have $x^\mu=(t, \mathbf{x})$. Using a dot to represent differentiation with respect to t, we can write the Lagrangian density (10.24) as follows:

$$\mathscr{L}=\tfrac{1}{2}[\dot\phi^2-(\nabla\phi)^2]-\tfrac{1}{2}m^2\phi^2-U(\phi). \tag{10.26}$$

As in (10.18), the field conjugate to ϕ, denoted by π, is as follows:

$$\partial\mathscr{L}/\partial\dot\phi=\dot\phi=\pi. \tag{10.27}$$

The Hamiltonian density is then defined by the following equations:

$$\mathscr{H}=\pi\dot\phi-\mathscr{L}=\tfrac{1}{2}\dot\phi^2+\tfrac{1}{2}(\nabla\phi)^2+\tfrac{1}{2}m^2\phi^2+U(\phi). \tag{10.28}$$

If the function $U(\phi)$ is positive definite, we see that the Hamiltonian density $\mathscr{H}$ given by (10.28) is positive definite, in keeping with the definition of the Hamiltonian as the energy. The Hamiltonian H is then given by integrating the Hamiltonian density $\mathscr{H}$ over all three-space:

$$H=\int\mathscr{H}\,\mathrm{d}^3x=\int[\tfrac{1}{2}\pi^2+\tfrac{1}{2}(\nabla\phi)^2+\tfrac{1}{2}m^2\phi^2+U(\phi)]\,\mathrm{d}^3x. \tag{10.29}$$

To proceed further, we consider the analogy with non-relativistic quantum mechanics of N particles with position and momenta given, respectively, by q_r, p_r, $r=1, \ldots, N$. Setting the modified Planck's constant $\hbar=1$, we write the commutation relations as follows (see (10.14)–(10.16)):

$$q_rp_s-p_sq_r=\mathrm{i}\delta_{rs},\quad r, s=1, \ldots, N, \tag{10.30}$$

where δ_{rs} is the Kronecker delta. It is well known that (10.30) implies the following representation for p_s: $p_s=-\mathrm{i}\partial/\partial q_s$. Consider now the following equal time commutation relation between the field ϕ and the conjugate field π:

$$\phi(t, \mathbf{x})\ \pi(t, \mathbf{x}') - \pi(t, \mathbf{x}')\ \phi(t, \mathbf{x}) = i\delta(\mathbf{x} - \mathbf{x}'), \tag{10.31}$$

where $\delta(\mathbf{x} - \mathbf{x}')$ is the three-dimensional Dirac delta function. In analogy with the above non-relativistic case, it can be shown that (10.31) implies that the conjugate field π has the following representation:

$$\pi(t, \mathbf{x}) = -\mathrm{i}\delta/\delta\phi(t, \mathbf{x}), \tag{10.32}$$

where the right hand side is the functional derivative with respect to $\phi(t, \mathbf{x})$. Since we shall be considering all functional derivatives at a particular time t, we will usually suppress the time and write the derivative thus: $\delta/\delta\phi(\mathbf{x})$. Just as the wave function in non-relativistic quantum mechanics is a function of the q_r and the time t, so for quantum fields the wave function is a functional of the field ϕ and a function of the time t, written thus: $\Psi[\phi, t]$. Schrödinger's equation is then a functional differential equation given by (setting $\hbar = 1$ in (10.20)):

$$H\Psi[\phi, t] = \mathrm{i}\partial\Psi/\partial t, \tag{10.33}$$

where H is the operator derived from (10.29) by replacing π by the right hand side of (10.32):

$$H = \int \{-\tfrac{1}{2}[\delta/\delta\phi(\mathbf{x})]^2 + \tfrac{1}{2}(\nabla\phi)^2 + \tfrac{1}{2}m^2\phi^2 + U(\phi)\}\ \mathrm{d}^3x. \tag{10.34}$$

As in quantum mechanics, we can consider stationary states which are of the form

$$H\Psi[\phi, t] = \Psi[\phi]\exp(-\mathrm{i}Et), \tag{10.35}$$

so that (10.33) reduces to

$$H\Psi[\phi] = E\Psi[\phi]. \tag{10.36}$$

Let us note the structure of (10.36). The functional $\Psi[\phi]$ itself is in general independent of $\mathbf{x}$ and is a pure (complex) number. However, the expression $\delta\Psi/\delta\phi(\mathbf{x})$ or $\delta^2\Psi/(\delta\phi(\mathbf{x}))^2$ *is* dependent on $\mathbf{x}$. On the left hand side, after integration over the spatial volume the integral becomes independent of $\mathbf{x}$, i.e., it becomes a pure number, as is the right hand side. At this stage we have mechanically extended the Schrödinger equation in the quantum mechanical case to the case of the quantum field in a simple-minded manner. There are certain subtleties, some of which will emerge later.

As regards the physical interpretation of the wave functional Ψ, it is similar to that of the wave function in quantum mechanics. Just as $|\psi(t, q)|^2\ \mathrm{d}q$ gives the probability of finding the quantum mechanical particle between coordinate values q and $q + \mathrm{d}q$, so $|\Psi(\phi, t)|^2\ \delta\phi$ gives the

probability of the field configuration having values between ϕ and $\phi+\delta\phi$ at time t. The function ϕ here plays the role of the coordinate q. This definition needs to be made more precise, such as considering the measure of the space of functions, etc.

10.4 A functional differential equation

In this section we give a simple example of a functional differential equation and its solution, to prepare the reader for the more complicated functional differential equation to be encountered. We have already given a brief discussion of functionals. For later convenience we use q, q', instead of x, x', etc.

A useful example is the functional derivative of an exponential, as follows:

$$G[A]=\exp(F[A]), \tag{10.37}$$

$$\delta G/\delta A(q')=\exp(F[A])(\delta F/\delta A(q')). \tag{10.38}$$

Equation (10.38) can be established by expanding $\exp(F)$ in (10.37) as a power series in F and using the relations

$$[\delta/\delta A(q')](F^n)=(nF^{n-1})(\delta F/\delta A(q')). \tag{10.39}$$

The foregoing formulae are sufficient to determine the functional derivatives of all functionals of interest. As a final example, we consider the following case, which can be treated using the foregoing relations and which will lead to a typical functional differential equation and its solution. We evaluate the functional derivatives of the following functional:

$$W[A]=\exp\left[b\int g(q',q'')\,A(q')\,A(q'')\,\mathrm{d}q'\,\mathrm{d}q''\right]\equiv\exp Z[A], \tag{10.40}$$

where $g(q', q'')$ is some fixed function of q' and q'', and b is a constant. From (10.38) we see that

$$\delta W/\delta A(q)=(\delta Z/\delta A(q))W. \tag{10.41}$$

We have

$$\begin{aligned}\delta Z/\delta A(q)&=b\int g(q',q'')\,A(q')\,(\delta A(q'')/\delta A(q))\\&\quad+(\delta A(q')/\delta A(q))\,A(q'')\,\mathrm{d}q'\,\mathrm{d}q''\\&=b\int g(q',q'')\,(A(q')\,\delta(q-q'')+\delta(q-q')\,A(q''))\,\mathrm{d}q'\,\mathrm{d}q''\\&=b\int g(q',q)\,A(q')\,\mathrm{d}q'+b\int g(q,q'')\,A(q'')\,\mathrm{d}q''\\&=b\int(g(q',q)+g(q,q'))\,A(q')\,\mathrm{d}q'\\&=2b\int g(q,q')\,A(q')\,\mathrm{d}q'.\end{aligned} \tag{10.42}$$

In the last step we have made the assumption that the function $g(q, q')$ is symmetric:

$$g(q, q') = g(q', q), \tag{10.43}$$

a relation which need not necessarily hold. Substituting in (10.41) we get

$$\delta W/\delta A(q) = (2b \int g(q, q')\, A(q')\, \mathrm{d}q')\, W. \tag{10.44}$$

Next we take the second functional derivative of W, for which we first take the functional derivative of the right hand side of (10.44) with respect to $A(\bar{q})$ rather than $A(q)$:

$$\begin{aligned}\delta^2 W/\delta A(q)\delta A(\bar{q}) &= 2b(\delta W/\delta A(\bar{q})) \int g(q, q')\, A(q')\, \mathrm{d}q' \\ &\quad + 2bW(\delta/\delta A(\bar{q})) \int g(q, q')\, A(q')\, \mathrm{d}q' \\ &= 4b^2 W \int\int g(q, q')\, g(\bar{q}, q'')\, A(q')\, A(q'')\, \mathrm{d}q'\, \mathrm{d}q'' \\ &\quad + 2bWg(q, \bar{q}),\end{aligned} \tag{10.45}$$

where we have used (10.42) and (10.44). We now set $q = \bar{q}$ to get

$$\delta^2 W/(\delta A(q))^2 = 4b^2 W[\int g(q, q')\, A(q')\, \mathrm{d}q']^2 + 2bW\, g(q, q). \tag{10.46}$$

This expression assumes that $g(q, q)$ is well defined, which may not be the case. For example, if $g(q, q') = (q - q')^{-1}$, then $g(q, q)$ is not defined. In this case one may have to use a convergence factor or use some other suitable limiting procedure.

Consider now the following functional differential equation of second order for the unknown functional $W'[A]$ of the function $A(q)$:

$$\int \{(\delta/\delta A(q))^2 + k\, (\int g(q, q')\, A(q')\, \mathrm{d}q')^2\} \mathrm{d}q\; W'[A] = EW'[A], \tag{10.47}$$

where k, E are constants and $g(q, q')$, as before, is a fixed function of q, q', which is symmetric in q, q'. From (10.40) and (10.46) it is clear that the functional $W[A]$ given by (10.40) is a solution of this equation if we identify the constants k, E as follows:

$$k = 4b^2, \quad E = 2b \int g(q, q)\, \mathrm{d}q. \tag{10.48}$$

The functional differential equation (10.47) is somewhat analogous to but simpler than the Schrödinger functional equation (10.48) for the scalar field. Equation (10.47) and its solution (10.40) therefore give some idea of the kind of solution we can expect, at least in the simple cases, and the manner in which the solution satisfies the functional differential equation.

10.5 Solution for a scalar field

The Schrödinger equation (10.34) reduces to that for a free scalar field if in the expression (10.34) for H we set $U=0$. We will simply state the solution for the ground state. The fact that it is a solution can be verified by the methods above. We consider first the massless case with $m=0$. In this case the ground state is given as follows:

$$\Psi_0[\phi]=N\exp\{-k\int f(\mathbf{x}',\mathbf{x}'')\,\partial_a^{(\mathbf{x}')}\phi(\mathbf{x}')\,\partial_a^{(\mathbf{x}'')}\phi(\mathbf{x}'')\,\mathrm{d}^3x'\,\mathrm{d}^3x''\},$$
$$\partial_a^{(\mathbf{x}')}=\partial/\partial x'^a,\ \text{etc.}, \tag{10.49}$$

where N, k are suitable constants and $f(\mathbf{x}',\mathbf{x}'')$ is a suitable weight function, defined below. Latin indices in (10.49) and other cases take values 1, 2, 3 in three-dimensional cases and values 1, 2 in cases where we have two spatial dimensions.

The massive case with non-zero m has the following ground state:

$$\Psi[\phi]=F[\phi]\Psi_0[\phi], \tag{10.50}$$

where $\Psi_0[\phi]$ is given by (10.49) and the functional $F[\phi]$ is defined as follows:

$$F[\phi]=\exp(J[\phi]),\quad J[\phi]=-k'\int j(\mathbf{x}',\mathbf{x}'')\,\phi(\mathbf{x}')\,\phi(\mathbf{x}'')\,\mathrm{d}^3x'\,\mathrm{d}^3x'', \tag{10.51}$$

where k' is a constant and the function $j(\mathbf{x}',\mathbf{x}'')$ is symmetric: $j(\mathbf{x}',\mathbf{x}'')=j(\mathbf{x}'',\mathbf{x}')$. The functions $f(\mathbf{x}',\mathbf{x}'')$ and $j(\mathbf{x}',\mathbf{x}'')$ are defined through their Fourier transforms $\hat{f}(\mathbf{p})$, $\hat{\jmath}(\mathbf{p})$:

$$f(\mathbf{x}',\mathbf{x}'')=(2\pi)^{-3}\int\hat{f}\,(\mathbf{p})\,\mathrm{e}^{-\mathrm{i}\mathbf{p}\cdot(\mathbf{x}'-\mathbf{x}'')}\,\mathrm{d}^3p, \tag{10.52a}$$

$$j(\mathbf{x}',\mathbf{x}'')=(2\pi)^{-3}\int\hat{\jmath}\,(\mathbf{p})\,\mathrm{e}^{-\mathrm{i}\mathbf{p}\cdot(\mathbf{x}'-\mathbf{x}'')}\,\mathrm{d}^3p. \tag{10.52b}$$

The functions $\hat{f}\,(\mathbf{p})$ and $\hat{\jmath}\,(\mathbf{p})$ are given as follows ($\hat{k}$ is a constant):

$$\hat{f}\,(\mathbf{p})=\hat{k}(\mathbf{p}^2)^{-1/2},\ \hat{\jmath}\,(\mathbf{p})=[-k\hat{k}(\mathbf{p}^2)^{1/2}+\tfrac{1}{2}(4k^2\hat{k}^2\mathbf{p}^2+m^2)^{1/2}]/k', \tag{10.53}$$

which gives a complete solution of the Schrödinger functional equation for the free massive scalar field.

10.6 The free electromagnetic field

The Lagrangian density for the free electromagnetic field can be written as follows:

$$\mathscr{L}=-(1/4)\,F_{\mu\nu}F^{\mu\nu},\quad F_{\mu\nu}=\partial_\mu A_\nu-\partial_\nu A_\mu, \tag{10.54}$$

A_μ being the electromagnetic four-potential; it is given in its contravariant form as $A^\mu = (A^0, \mathbf{A})$, where A^0 is the electric potential and $\mathbf{A}$ is the three-vector potential. The Lagrangian (10.54) is invariant under the gauge transformation

$$A_\mu \to A'_\mu = A_\mu + \partial_\mu V, \tag{10.55}$$

where V is any arbitrary function of space-time. It is well known that one can use this freedom to introduce the *temporal gauge* in which the time component of the four-vector potential vanishes: $A_0 = 0 = A^0$. The Lagrangian density (10.54) then reduces to the following expression:

$$\mathscr{L} = \tfrac{1}{2}\dot{A}_n\dot{A}_n - (1/4)F_{ab}F_{ab}, \tag{10.56}$$

where Latin indices range over values 1, 2, 3 and a dot, as before, denotes differentiation with respect to t. The momentum canonically conjugate to A_n is given by P_n, where

$$P_n = \partial\mathscr{L}/\partial\dot{A}_n = \dot{A}_n, \tag{10.57}$$

so that the Hamiltonian density is

$$\mathscr{H} = P_n\dot{A}_n - \mathscr{L} = \tfrac{1}{2}P_nP_n + (1/4)\ F_{ab}F_{ab}. \tag{10.58}$$

Since contravariant and covariant spatial indices differ by a sign only, in (10.56) and (10.58) it does not matter if we use $F_{ab}F_{ab}$ or $F_{ab}F^{ab}$, as they are equal. Since P_n is canonically conjugate to A_n, when we quantize the theory $P_n(\mathbf{x})$ becomes the operator $-\mathrm{i}\delta/\delta A_n(\mathbf{x})$, in a manner similar to the case of the scalar field, where we have again suppressed the t-dependence. Following steps similar to the case of the scalar field, the Schrödinger equation for stationary states in this case can be written as

$$H\Psi[A] = \tfrac{1}{2}\int\{-\delta^2/[\delta A_n(\mathbf{x})]^2 + \tfrac{1}{2}F_{ab}(\mathbf{x})F_{ab}(\mathbf{x})\}\ \Psi[A]\ \mathrm{d}^3x = E\Psi[A]. \tag{10.59}$$

Here $\Psi[A]$ is the stationary state which is a functional of the three components of $\mathbf{A}$. There is a significant difference between the cases of the scalar and the electromagnetic field. The condition $A_0 = 0$ still leaves an arbitrariness in the remaining components A_1, A_2, A_3 because we can still carry out a purely spatial gauge transformation

$$A_n \to A'_n = A_n + \partial_n h, \tag{10.60}$$

where the function h is independent of the time and a function of $\mathbf{x}$ only. This transformation will leave the modified Lagrangian density (10.56) invariant. The three functions $\mathbf{A}$ in (10.59) are therefore not independent

in this sense. It can be shown that this invariance implies that in addition to the Schrödinger equation (10.59) the wave functional must satisfy the gauge constraint or Gauss's relation, as follows:

$$\partial_n^{(\mathbf{x})}(\delta\Psi/\delta A_n(\mathbf{x})) = 0. \tag{10.61}$$

The ground state solution of (10.59) has been considered by various authors and is well known. It is given as follows:

$$\Psi_0[A] = N' \exp\{-k'' \int F_{ab}(\mathbf{x}')\, F_{ab}(\mathbf{x}'')\, g(\mathbf{x}', \mathbf{x}'')\, \mathrm{d}^3x'\, \mathrm{d}^3x''\}, \tag{10.62}$$

where N', k'' are constants and $g(\mathbf{x}', \mathbf{x}'')$ is a suitable weight function [independent of the $A_n(\mathbf{x})$]. In fact, $g(\mathbf{x}', \mathbf{x}'')$ turns out to be proportional to the weight function $f(\mathbf{x}', \mathbf{x}'')$ considered in the scalar case, and can be identified with the latter. The case of the electromagnetic field considered in this section, although very simple and well known, provides a useful background for the much more complicated case of the Yang–Mills field, or the Wheeler–De Witt equation. The case of linearized gravity considered by Hartle (1984) is similar to this case.

10.7 The Wheeler–De Witt equation

In quantum gravity one can derive an equation similar to the Schrödinger equation (10.36), which is known as the Wheeler–De Witt equation. This equation is best derived from the path integral formalism, which we will consider in the next section. We shall not give the derivation here but only write down the equation itself and give a brief description of it.

As mentioned earlier, there are many subtleties which we will not consider. One of these is the manner in which space-time should be sliced to give a series of appropriate three-geometries, in which there remains the problem of dealing with the freedom of carrying out spatial transformations. One of the conditions that go into the derivation of the Wheeler–De Witt equation (that given, for example, by Hartle and Hawking (1983)), is that the universe should be closed, so that the space-like sections are compact. We use units with $\hbar = c = 1$ and introduce coordinates so that the space-like hypersurfaces are $t =$ constant and the metric is written as follows:

$$\mathrm{d}s^2 = (N^2 - N_i N^i)\, \mathrm{d}t^2 - 2N_i\, \mathrm{d}x^i \mathrm{d}t - h_{ij}\, \mathrm{d}x^i \mathrm{d}x^j, \quad i, j = 1, 2, 3. \tag{10.63}$$

N, N^i are functions of space-time with $N_i = h_{ij} N^j$. K_{ij} is the extrinsic curvature of the three-surface $t =$ constant given as follows:

$$K_{ij} = -n_{i;j}, \tag{10.64}$$

where n^i is the spatial part of the unit normal to the hypersurface, t = constant and the semi-colon denotes covariant derivative as in (2.6b). Equation (10.64) can be written as follows (see Appendix A14):

$$K_{ij} = \tfrac{1}{2}N^{-1}(\dot{h}_{ij} - \nabla_i N_j - \nabla_j N_i), \tag{10.65}$$

where ∇_j denotes covariant derivative with respect to the three-metric h_{ij}. The momentum canonically conjugate to h_{ij} is given as follows in terms of K_{ij} and its trace $K = K^i_i$:

$$\pi_{ij} = -h^{1/2}(K_{ij} - h_{ij}K), \tag{10.66}$$

where h is the determinant of the metric h_{ij}. The wave functional in this case is a functional $\Psi[h_{ij}]$ of the three-metric h_{ij} and is related to the probability of finding the space-like hypersurface with the given three-metric. One can find an expression which is equivalent to the Hamiltonian, and if one replaces the π^{ij} with the operator $-\mathrm{i}\delta/\delta h_{ij}$ in it, one gets the Wheeler–De Witt equation

$$\left(-G_{ijkm}\frac{\delta^2}{\delta h_{ij}\delta h_{km}} - {}^3Rh^{1/2} + 2\Lambda h^{1/2}\right)\Psi[h_{ij}] = 0. \tag{10.67}$$

Here

$$G_{ijkm} = \tfrac{1}{2}h^{-1/2}(h_{ik}h_{jm} + h_{im}h_{jk} - h_{ij}h_{km}), \tag{10.68}$$

${}^3R$ is the scalar curvature for the three-metric, and Λ the cosmological constant. Equation (10.67) corresponds to the stationary form of Schrödinger's equation given by $H\Psi = E\Psi$. The tensor G_{ijkm} is, in fact, the metric in the 'superspace', which is the space of all three-geometries. In (10.67) we have also ignored the matter fields, for which one would get additional terms. The freedom to carry out spatial transformations of the three-metric gives additional constraints which the wave function must satisfy – these are familiar in gauge theories in the so-called Gauss relations. This completes our brief discussion of the Wheeler–De Witt equation. We will now consider the equivalent path integral approach, for which we first give a brief account of path integrals.

10.8 Path integrals

In recent years path integrals have been used increasingly in the formulation of gauge theories and other aspects of physics. The originator of the method of path integrals was Feynman (1948) (see also Feynman and Hibbs, 1965). There are many introductory accounts of path integrals (see,

for example, Taylor, 1976). We will give the bare essentials here (see Narlikar and Padmanabhan, 1983).

A convenient way of introducing path integrals is to compare the formulation of the equations of motion of a free particle in classical mechanics and quantum mechanics as expressed in terms of path integrals. If the position vector of the particle is $\mathbf{r} = (x, y, z)$, its equation of motion is given by

$$m\ddot{\mathbf{r}} = 0. \tag{10.69}$$

Suppose the particle is at $\mathbf{r}_\mathrm{i}$ at time t_i (the initial time) and at $\mathbf{r}_\mathrm{f}$ at time t_f (the final time). It is easy to see that in the intervening period $t_\mathrm{i} < t < t_\mathrm{f}$ the position vector is given by

$$\mathbf{r}(t) = \mathbf{r}_\mathrm{i} + \left(\frac{t - t_\mathrm{i}}{t_\mathrm{f} - t_\mathrm{i}}\right)(\mathbf{r}_\mathrm{f} - \mathbf{r}_\mathrm{i}) \equiv \bar{\mathbf{r}}(t). \tag{10.70}$$

Quantum mechanically, if the particle is at $\mathbf{r}_\mathrm{i}$ at time t_i, one can only give a *probability amplitude* for finding the particle at $\mathbf{r}_\mathrm{f}$ at time t_f; this is given as follows:

$$K(\mathbf{r}_\mathrm{f}, t_\mathrm{f}; \mathbf{r}_\mathrm{i}, t_\mathrm{i}) = [m/2\pi\mathrm{i}\hbar(t_\mathrm{f} - t_\mathrm{i})]^{3/2} \exp[\mathrm{i}m(\mathbf{r}_\mathrm{f} - \mathbf{r}_\mathrm{i})^2/2\hbar(t_\mathrm{f} - t_\mathrm{i})]. \tag{10.71}$$

The connection between (10.70) and (10.71) is established by saying that classically the particle follows the definite path $\bar{\mathbf{r}}(t)$ given by (10.70) whereas quantum mechanically the particle can take any path that is allowed by causality; there is a certain amplitude associated with each path $\mathbf{r}(t)$ and to get the complete amplitude to find the particle at $\mathbf{r}_\mathrm{f}$ at time t_f one has to sum over all paths weighted by the amplitude for the path (see Fig. 10.2). The amplitude associated with the path $\mathbf{r}(t)$ is given by $\exp[(\mathrm{i}/\hbar)S]$ where S is the classical action assoociated with the path, given by (10.72) (with $\mathbf{r}(t)$ instead of q; that is, instead of the one coordinate q we have three coordinates $\mathbf{r}$). The sum is then taken over all paths; this sum is a kind of integral over all functions $\mathbf{r}(t)$, which can be defined by a limiting procedure in which the time interval $[t_\mathrm{i}, t_\mathrm{f}]$ is divided into n equal parts so that the weight function becomes a function of the variables $\mathbf{r}(t_\mathrm{m})$, where t_m is a typical instant at which the division of the interval $[t_\mathrm{i}, t_\mathrm{f}]$ occurs. One then integrates over the $\mathbf{r}_\mathrm{m} \equiv \mathbf{r}(t_\mathrm{m})$, and takes the limit as n tends to infinity to get the complete amplitude to arrive at $\mathbf{r}_\mathrm{f}$ at time t_f (see, for example, Feynman and Hibbs, 1965, for details). For example, if one starts with the classical action which gives rise to (10.69) namely,

$$S = \int_{t_\mathrm{i}}^{t_\mathrm{f}} \tfrac{1}{2} m\dot{\mathbf{r}}^2 \, \mathrm{d}t, \tag{10.72}$$

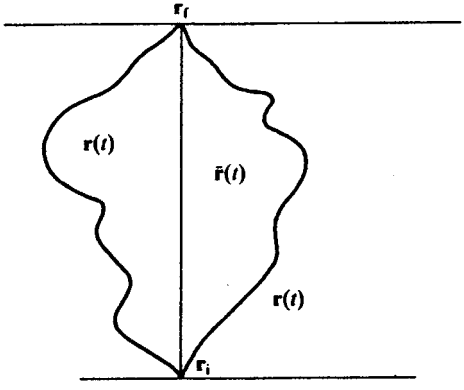


Fig. 10.2. Classically the particle follows the path $\bar{\mathbf{r}}(t)$; quantum mechanically the particle can follow any of the paths $\mathbf{r}(t)$, but each path is weighted by the amplitude $\exp(\mathrm{i}/\hbar)S$, where S is the classical action associated with the path $\mathbf{r}(t)$.

one arrives at the amplitude (10.71) by adopting this limiting procedure. Symbolically, this integral can be written as follows:

$$K(\mathbf{r}_f, t_f; \mathbf{r}_i, t_i) = \int \exp\{(\mathrm{i}/\hbar)S[\mathbf{r}(t)]\}\, \mathscr{D}\mathbf{r}(t). \tag{10.73}$$

Note that here S is not a function but a functional of $\mathbf{r}(t)$; for this reason this integral is also called a functional integral. Here the symbol $\mathscr{D}\mathbf{r}(t)$ means the integral is a sum over all functions $\mathbf{r}(t)$ in the sense explained in Feynman and Hibbs (1965).

By dividing every path at a certain instant t, one can derive from (10.73) the following relation:

$$K(\mathbf{r}_f, t_f; \mathbf{r}_i, t_i) = \int K(\mathbf{r}_f, t_f; \mathbf{r}, t) K(\mathbf{r}, t; \mathbf{r}_i, t_i)\, \mathrm{d}^3\mathbf{r}, \tag{10.74}$$

for any time $t_i \leqslant t \leqslant t_f$. Similarly, one can show that if the state of the particle at time t_i is represented by the wave function $\psi_i(\mathbf{r}_i, t_i)$, then its final wave function $\psi_f(\mathbf{r}_f, t_f)$ is given as follows:

$$\psi(\mathbf{r}_f, t_f) = \int K(\mathbf{r}_f, t_f; \mathbf{r}_i, t_i)\psi_i(\mathbf{r}_i, t_i)\, \mathrm{d}^3\mathbf{r}_i. \tag{10.75}$$

This relation can be verified explicitly for the free particle wave function $\psi(\mathbf{r}_i, t_i) \sim \exp[(\mathrm{i}/\hbar(Et_i - \mathbf{p}\cdot\mathbf{r}_i)]$, with a similar expression for $\psi(\mathbf{r}_f, t_f)$ where $E = p^2/2m$ if one uses the K given by (10.73). In fact this confirms that (10.71) is the corrrect free particle amplitude.

By analogy with the above case, one can consider the case of a spacetime geometry, in which Σ_i, Σ_f represent respectively an initial and a final

space-like hypersurface (see Fig. 10.1(b)), and one asks for the probability amplitude for a certain three-geometry ${}^{(3)}\mathscr{G}_{\mathrm{f}}$ on Σ_{f} given the three-geometry ${}^{(3)}\mathscr{G}_{\mathrm{i}}$ on Σ_{i}. In this case the classical 'path' is the solution given by Einstein's equations, but the contributions to the probability amplitude come from all four-geometries which are not necessarily solutions of Einstein's equations. Symbolically this can be represented by

$$K\{{}^{(3)}\mathscr{G}_{\mathrm{f}}, \Sigma_{\mathrm{f}}; {}^{(3)}\mathscr{G}_{\mathrm{i}}, \Sigma_{\mathrm{i}}\} = \int \exp\{(\mathrm{i}/\hbar)S[g]\}\,\mathscr{D}g, \tag{10.76}$$

in analogy with (10.73). Here $S[g]$ is the action for gravitation (see, e.g., (10.77) below) and the functional integration is over all four-geometries connecting Σ_{i} and Σ_{f}. There are, of course, many complexities hidden in (10.76). For example, one has to take into account that some four-geometries are simply transforms of each other. Presumably these can be taken into account by a generalization of the method of Faddeev and Popov (1967) which is used in Yang–Mills type gauge theories and which effectively amounts to dividing out an infinite gauge volume. Secondly, the actual evaluation of the path integral (10.76) for any given situation presents prohibitive problems. Nevertheless, the conceptual simplicity of (10.76) is striking. In the next section we will examine how some information can be extracted from (10.76) with some simplifying assumptions.

We end this section with some remarks about the classical limit of the path integral (10.73). Classical physics is valid when the action of the classical system is large compared to $\hbar$; note that S and $\hbar$ have the same dimensions so that $S/\hbar$ is a pure number. Thus for a classical system the phase of the exponential in (10.73) is large for most paths, so that a small variation in the path causes a relatively large variation in the phase with the result that, because of the oscillation of the exponential, the contributions from neighbouring paths cancel each other. The only paths which contribute substantially in this case are those for which the action does not vary much with the variation in the paths. These are given by paths near the one which gives $\delta S = 0$, which, of course, yields the classical path $\bar{\mathbf{r}}(t)$. Thus in the limit of vanishing $\hbar$ we get the classical path. The interesting thing is that the same argument applied to (10.76) yields the Einstein equations in the classical limit, these equations being given by $\delta S_{\mathrm{g}} = 0$, where S_{g} is given as follows, inserting c:

$$S_{\mathrm{g}} = (c^3/16\pi G)\int_{\mathscr{V}} R(-g)^{1/2}\,\mathrm{d}^4x, \tag{10.77}$$

where $\mathscr{V}$ is the space-time region under consideration. The scalar curvature R of the space-time has dimensions (length)$^{-2}$. Thus if we take

$R \sim L^{-2}$, where L is the characteristic length, and the four-volume $\mathscr{V}$ is of dimension L^4, we find the following estimate for the magnitude of S_g:

$$S_g = c^3 L^2 / 16\pi G. \tag{10.78}$$

Thus the action S_g becomes comparable to $\hbar$ if the linear size of the universe is (ignoring the numerical factor 16π)

$$L_P = (G\hbar/c^3)^{\frac{1}{2}} \sim 1.6 \times 10^{-33} \text{ cm}, \tag{10.79}$$

which is the so-called *Planck length*.

We note finally that the Schrödinger equation can be derived from (10.75) by making $t_f - t_i$ infinitesimally small.

10.9 Conformal fluctuations

We have seen that in the path integral (10.76) the sum involves space-times which do not necessarily satisfy Einstein's equations. In practice to include all such space-times is a formidable task. One simplification that has been tried is to consider only geometries which are conformal to the classical solutions, that is, solutions of Einstein's equations. Suppose that for a given action (10.77) we have a classical solution given by the metric

$$\mathrm{d}\bar{s}^2 = \bar{g}_{\mu\nu}\, \mathrm{d}x^\mu\, \mathrm{d}x^\nu, \tag{10.80}$$

for the region which lies between the space-like hypersurfaces Σ_i and Σ_f (see Fig. 10.1(*b*)). Non-classical paths also contribute to (10.76) but we consider only those paths which are conformally related to (10.80), that is, only those metrics which are of the following form:

$$\mathrm{d}s^2 = \Omega^2\, \mathrm{d}\bar{s}^2 = \Omega^2 \bar{g}_{\mu\nu}\, \mathrm{d}x^\mu\, \mathrm{d}x^\nu, \tag{10.81}$$

where Ω is an arbitrary function of space-time. Since Einstein's equations are not conformally invariant, except in the trivial case $\Omega = \text{constant}$, (10.81) represents a non-classical path between Σ_i and Σ_f. There are other ways of generating non-classical paths, but the merit of (10.81) is that null geodesics are conformally invariant, so the light cone structure of space-time is preserved by such paths. We write

$$\phi = \Omega - 1, \tag{10.82}$$

so that ϕ represents the conformal fluctuation around the classical path. We shall only give the results of the consideration of conformal paths, and refer to Narlikar and Padmanabhan (1983) for the details. We take the

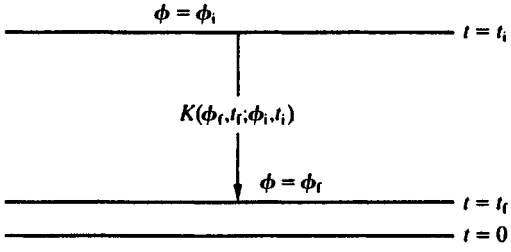


Fig. 10.3. For the conformal fluctuation of Friedman cosmologies we reverse the time and take the final time t_f to be earlier than the initial time t_i, with the former near the singularity at $t = 0$.

classical geometry to be that of Friedmann cosmologies, which we write as follows

$$d\bar{s}^2 = dt^2 - Q^2(t)[dr^2/(1-kr^2) + r^2(d\theta^2 + \sin^2\theta\, d\phi^2)]. \tag{10.83}$$

We consider the state of the universe at the initial epoch t_i to be given by a wave packet with spread Δ_i, as follows:

$$\psi_i(\phi, t_i) = (2\pi\Delta_i^2)^{-1/4} \exp(-\phi^2/4\Delta_i^2). \tag{10.84}$$

It is shown by Narlikar (1979) and by Narlikar and Padmanabhan (1983) that if conformal paths are taken into account, the wave packet (10.84) evolves to the one given by a similar expression to (10.84) except that Δ_i is replaced by Δ_f given as follows (see (10.75) and Fig. 10.3):

$$\Delta_f = (2\pi T/3VQ_iQ_f)[1 + (3V/2\pi T)\Delta_i^2 Q_i^2(1 + TQ_iH_i)^2]^{1/2}, \tag{10.85}$$

where V is the coordinate volume of the region under consideration, given by $r \leqslant r_b$, where r is the radial coordinate occurring in (10.83) and T, H_i are defined as follows:

$$T = \int_{t_f}^{t_i} du/Q(u), \quad H_i = \dot{Q}(t_i)/Q(t_i), \quad Q_i = Q(t_i),\ Q_f = Q(t_f). \tag{10.86}$$

The important thing to notice is that as t_f tends to zero, that is, as we approach the singularity, Δ_f goes as Q_f^{-1}, and so diverges. Thus it appears that in the limit of the classical singularity quantum conformal fluctuations diverge. Thus the classical solution, which can be regarded as the 'average' of the wave packet, is no longer reliable near the singularity. Narlikar and Padmanabhan (1983) further find indications that quantum conformal fluctations may prevent a space-time singularity and also may eliminate the appearance of a particle horizon.

There are many approximations involved in the above considerations and, hence, many uncertainties. It is not clear to what extent the claims made in the above paragraph are valid. The important thing to notice here, however, is that the formalism of this chapter seems to provide a handle with which these interesting questions can be meaningfully tackled. There is obviously a long way to go before definitive answers can be given to such questions. The above work has been generalized by Joshi and Narlikar (1986) to cases where the state of the universe is defined by wave functionals that are not necessarily wave packets, with similar results.

To end this section we consider as an illustration conformal perturbation of flat space-time. For this we first consider a transform of S_g given by (10.77) under the conformal transformation (10.81). Putting $c=\hbar=1$, S_g is transformed to S'_g given by the following expression:

$$S'_g=(16\pi G)^{-1}\int_{\mathcal{V}}(\Omega^2\bar{R}-6\Omega_\mu\Omega^\mu)(-\bar{g})^{1/2}\,\mathrm{d}^4x,\ \Omega_\mu=\frac{\partial}{\partial x^\mu}\Omega, \qquad (10.87)$$

where $\bar{R}$ is the scalar curvature derived from the metric $\bar{g}_{\mu\nu}$, and $\bar{g}$ is the determinant of this metric. If we now specialize the metric $\bar{g}_{\mu\nu}$ to that of flat space given by $\eta_{\mu\nu}$ with $-\eta_{00}=\eta_{11}=\eta_{22}=\eta_{33}=-1$, with $\eta_{\mu\nu}=0$ when $\mu\neq\nu$, the action S'_g reduces to the following one:

$$S_\Omega=-(3/8\pi G)\int\Omega_\mu\Omega^\mu\mathrm{d}^4x. \qquad (10.88)$$

We apply the formalism developed in (10.24)–(10.29). The Lagrangian density for (10.88) is given as follows:

$$\mathscr{L}=(-3/8\pi G)\Omega_\mu\Omega^\mu=(-3/8\pi G)(\dot{\Omega}^2-(\nabla\Omega)^2), \qquad (10.89)$$

so that the momentum density canonically conjugate to Ω is given by

$$\partial\mathscr{L}/\partial\dot{\Omega}=2k'\dot{\Omega}=\pi,\quad k'\equiv(-3/8\pi G). \qquad (10.90)$$

The Hamiltonian density is given as follows:

$$\mathscr{H}=\pi\dot{\Omega}-\mathscr{L}=k'[\dot{\Omega}^2+(\nabla\Omega)^2]=(4k')^{-1}\pi^2\quad+k'(\nabla\Omega)^2. \qquad (10.91)$$

The corresponding Schrödinger equation is (replacing π by $-\mathrm{i}\delta/\delta\Omega$):

$$(4k')^{-1}\int[-(\delta/\delta\Omega)^2+4k'^2(\nabla\Omega)^2]\,\mathrm{d}^3x\,\Psi[\Omega]=E\Psi[\Omega]. \qquad (10.92)$$

This equation is similar to the one that obtains in quantum electrodynamics of the pure electromagnetic field, for which the solution is well known (see, for example, Rossi and Testa, 1984; Hartle, 1984; Islam, 1989; also Feynman and Hibbs, 1965. The ground state solution of (10.92) can be written as follows (the derivation is similar to that of (10.49), (10.62)):

$$\Psi[\Omega] = N \exp\left[\left(-\frac{3}{8\pi^3 L_{\mathrm{P}}^2}\right)\int\int \frac{\Delta_x \Omega \cdot \nabla_y \Omega}{(\mathbf{x}-\mathbf{y})^2}\,\mathrm{d}^3x\,\mathrm{d}^3y\right], \tag{10.93}$$

which gives the probability amplitude for detecting a conformal factor in flat space. This expression implies that large deviations from flat space can occur at Planck length scales L_{P}. This is usually referred to as the 'foam' structure of space-time.

10.10 Further remarks about quantum cosmology

We end this chapter by mentioning some further developments. One significant one is the proposal for the wave function of the 'ground state' of the universe, put forward by Hartle and Hawking (1983), which we describe here briefly. An interesting aspect of any quantum mechanical theory is the ground state or the state of minimum excitation. In terms of path integrals, the ground state at $t=0$ can be defined by

$$\psi_0(x, 0) = N\int \exp\{-I[x(\tau)]\}\,\mathscr{D}x(\tau); \tag{10.94}$$

where the time integral in the action S has been transformed by $t \to -\mathrm{i}\tau$, to make the path integral well defined (this does not, in general, affect its value) and $\mathrm{i}S$ has been replaced by $-I$. The function $x(\tau)$ represents all paths which end at $x(0)=x$ at $t=\tau=0$. (A proof of (10.94) can be found in Hartle and Hawking (1983).)

In the case of closed universes, which Hartle and Hawking consider, it is not appropriate to define the ground state as the state of lowest energy, as there exists no natural definition of energy for a closed universe. In fact, the total energy of a closed universe may be zero – the gravitation and matter energies cancelling each other. It might be reasonable, however, to define a state of minimum excitation corresponding classically to a geometry of high symmetry. In analogy with (10.94) Hartle and Hawking propose the following expression as the ground state wave function of a closed universe:

$$\psi_0[h_{ij}] = N\int \exp(-I_{\mathrm{E}}[g])\,\mathscr{D}g, \tag{10.95}$$

where I_{E} is the Euclidean action for gravity (obtained by carrying out the transformation $t \to -\mathrm{i}\tau$ in S_{g} given by (10.77)) and including the cosmological constant Λ. They are able to work out the path integral using certain simplifying assumptions, and find that the ground state corresponds to de Sitter space in the classical limit. They also find excited states which yield universes which start from zero volume, reach a maximum and collapse,

but which also have a non-zero (but small) probability of tunnelling through a potential barrier to a de Sitter type of continued expansion.

We have glossed over several complexities earlier in the chapter. One of these is the problem of 'operator ordering' in (10.67) where a simple ordering between π_{ij} and h_{ij} has been used. Another possibility for the first term in (10.67) would have been, for example,

$$(2h^{1/2})^{-1}(\delta/\delta h_{ij})h^{1/2}G_{ijkm}(\delta/\delta h_{km}). \tag{10.96}$$

A term q^2p, for example, in the classical Hamiltonian, can become q^2p, qpq, or pq^2 and one has to use other considerations to decide which is the correct one, as quantum mechanically these are apparently distinct possibilities, since here q, p are non-commuting.

Coleman and Banks (see Schwarzschild, 1989), and references therein) have considered a modification of the Hartle–Hawking formalism given by (10.94) and (10.95) in which the path integrals are not only over the entire history of the present universe, but also over the full manifold of all universes connected by wormholes (see, for example, Misner, Thorne and Wheeler, 1973, p. 1200). In the resulting analysis they find an explanation of the vanishing of the cosmological constant (see also Weiss, 1989).

We have attempted to provide here the bare minimum of the subject of quantum cosmology. It is hoped that this will enable the reader to follow the more specialized material in the papers cited here (see particularly Hartle, 1986, and the papers cited there).

11
The distant future of the universe

11.1 Introduction

In the previous chapters we have considered in some detail the 'standard' model of the universe. It is pertinent to ask what the prediction of the standard model is for the distant future of the universe. The future of the universe has been the subject of much speculation, in one form or another, from time immemorial. It is only in the last few decades that enough progress has been achieved in cosmology to study this question scientifically. In this chapter we shall attempt to provide an account of – or at any rate limit the possibilities for – the distant future of the universe, on the basis of the present state of knowledge. We refer the reader to Rees (1969), Davies (1973), Islam (1977, 1979a,b, 1983a,b), Barrow and Tipler (1978) and Dyson (1979) for more material on this topic. This chapter is based mostly on the papers by Islam and Dyson.

The distant future of the universe is dramatically different depending on whether it expands forever, or it stops expanding at some future time and recollapses. In the earlier chapters we have considered in detail the conditions under which these possibilities are likely to arise. As galaxies are the basic constituents of the universe, to examine the distant future of the universe we must consider the long term evolution of a typical galaxy. We will first assume that we are in an open universe, or, at any rate, that an indefinite time in the future is available. It is worth noting that by taking the mass density of the universe to be above but sufficiently close to the critical density, we can get models of the universe which have a finite but arbitrarily long life-time.

11.2 Three ways for a star to die

In any amount of matter there is a tendency for the matter to collapse towards the centre of mass due to the gravitational attraction of different

parts for each other. In a star this inward force is balanced by the release of energy during nuclear burning in which hydrogen is converted into helium and helium into heavier nuclei. At this stage the material of the star can be approximated by an ideal gas, in which the pressure p is related to its temperature T and number density n by the relation:

$$p = nkT, \tag{11.1}$$

where k is Boltzmann's constant (there should not be any confusion with the k used in the Robertson–Walker metric). As the star loses energy and its temperature decreases, this thermal energy, after a few billion years, is insufficient to balance the inward force of gravity. The star contracts and becomes more dense so that the electrons are eventually stripped off the atoms and run about freely in the material of the star. They then exert a Fermi pressure due to the Pauli exclusion principle. When the density is about 5×10^6 g cm^{-3} this electron degeneracy pressure is given by, restoring c (Chandrasekhar, 1939):

$$p \sim hcn^{4/3}. \tag{11.2}$$

At lower densities p is proportional to $n^{5/3}$.

For a spherically symmetric star, p and the mass density ρ satisfy the equation of hydrostatic equilibrium at radius r:

$$\mathrm{d}p/\mathrm{d}r = -G[m(r)/r^2]\rho, \tag{11.3}$$

where $m(r)$ is the mass inside radius r. One can show that in order to support itself against collapse the pressure p_c at the centre must be

$$p_c \sim GM^{2/3}\rho^{4/3}, \tag{11.4}$$

where M is the total mass of the star. Thus the pressure available at high densities (11.2) and the pressure needed for support have the same dependence on n (since ρ is proportional to n). It can be shown that when M is less than about 1.4 times the mass of the Sun, the electron degeneracy pressure can permanently halt collapse (Chandrasekhar, 1935, 1939) and one gets white dwarfs whose size is roughly that of the Earth. These eventually become cold and stop radiating altogether to become what are sometimes called 'black dwarfs'. The nuclei in these stars are mostly those of iron, since the latter has the most stable nucleus.

When the mass of the star is greater than 1.4 solar masses, or if there is a sudden inward pressure due to an explosion of the outer layers, the electron degeneracy pressure is insufficient to balance gravity. The star continues to collapse and becomes more dense until the electrons are squeezed

into the protons of the nuclei to become neutrons and different nuclei coalesce until the star becomes a giant nucleus – a neutron star. If the mass of the star is less than a certain critical mass M_c (this is about 2–3 solar masses) the neutron degeneracy pressure and the forces of nuclear interactions are sufficient to balance gravity. To find M_c one must appeal to general relativity, since Newtonian theory is inadequate for the strong fields generated by the neutron stars. For the latter (11.13) is replaced by

$$\mathrm{d}p/\mathrm{d}r = G(\rho + p/c^2)[m(r) + 4\pi r^3 p/c^2]/\{r[r - 2Gm(r)/c^2]\}. \qquad (11.5)$$

Equation (11.5) implies that more pressure is needed to support a star for strong fields than is implied by Newtonian theory. Neutron stars are the pulsars, discovered in 1967, of which more than six hundred have been found since the original discovery (Hewish *et al.*, 1968).

When the mass of the star is greater than M_c after shedding any mass, even neutron degeneracy pressure and the forces of nuclear interactions are insufficient to halt the collapse. In this case there is no known force which can halt the collapse and it is assumed that the star continues to collapse until it gets literally to a point – into a space-time singularity akin to the space-time singularity of the very early universe, about the nature of which, as seen earlier in this book, there is a great deal of uncertainty. This collapse results in a black hole which is a spherical region of radius $2GM/c^2$, where M is the mass of the star. If M is ten times the solar mass then this radius – the Schwarzschild radius – is about 20–30 km. The surface of the sphere of the Schwarzschild radius is called the horizon and the spherical region is called a black hole because once the star collapses to within this region nothing – not even light – can escape. There may, of course, be radiation from infalling matter just before the matter enters the region. The black hole may be detected by such radiation and also by its gravitational influence on nearby stars, etc. (see, for example, Thorne, 1974).

The above three final states, namely, those of black dwarf, neutron star and black hole, occur for masses not too small compared to the mass of the Sun. For smaller bodies such as the Earth and the Moon or a piece of rock, gravity can be balanced indefinitely by the ordinary pressure that matter exerts in resisting being compressed.

11.3 Galactic and supergalactic black holes

Consider the fate of a typical galaxy assuming we have an indefinite period ahead. All stars will ultimately be reduced to black dwarfs, neutron stars

or black holes. As the galaxy will be losing energy by radiation all the time, including the thermal energy of any hot interstellar gas, given sufficient time the galaxy will eventually consist of a gravitationally bound system of black holes, neutron stars, black dwarfs and cold interstellar matter in the form of planets, asteroids, meteorites, dust, etc. From the average energy and luminosity of a typical galaxy one can deduce that the time scale to arrive at this state will be anything between 10^{11} and 10^{14} years.

This situation will continue for thousands of billions of years without any significant changes within galaxies, but galaxies which are not in the same cluster will continue to recede from each other. The next significant change in a galaxy will take much longer than earlier changes such as stars becoming black holes, etc. The stars (henceforth by 'stars' we mean black dwarfs, neutron stars or stellar size black holes) in the galaxy will eventually tend to form a dense central core with an envelope of low density. The long term evolution of such a system is very difficult to predict accurately (see, for example, Saslaw, 1973, and Saslaw, Valtonen and Aarseth, 1974). Some stars, if they are involved in close three-body or many-body encounters, may be thrown out of the galaxy altogether. Such encounters are relatively rare in time scales of a few billion years. The time scales over which such processes dominate can be worked out as follows (Dyson, 1979). If a galaxy consists of N stars of mass M in a volume of radius R, their root-mean-square velocity will be of the order

$$v=(GNM/R)^{1/2}. \tag{11.6}$$

The cross-section for close collision is

$$\sigma=(GM/v^2)^2=(R/N)^2, \tag{11.7}$$

and the average time spent by a star between two collisions is

$$t_{\text{av}}=(\rho v\sigma)^{-1}=(NR^3/GM)^{1/2}, \tag{11.8}$$

where ρ is the density of stars in space. For a typical galaxy $N=10^{11}$, $R=3\times10^{17}$ km, so

$$t_{\text{av}}=10^{19} \text{ years}. \tag{11.9}$$

Dynamical relaxation of the galaxy takes about 10^{18} years. The combined effect of close collisions and dynamical relaxation is to produce a dense central core which eventually collapses to a single black hole, while stars from the outer regions evaporate in a time scale of that given by (11.9). The number of stars that will escape is very difficult to determine; perhaps 99%. Thus in about 10^{20} years or somewhat longer the original galaxy will

be reduced to a single 'galactic' black hole of about 10^9 solar masses, while stray stars and other small pieces of matter thrown out of the galaxy will be wandering singly in the intergalactic space.

It is likely that a cluster of galaxies will continue to be gravitationally bound as the expansion of the universe proceeds. Through long term dynamical evolution as described above the cluster will also eventually reduce to a single 'supergalactic' black hole of about 10^{11} or 10^{12} solar masses, a large fraction of the stars having evaporated.

This process of the transformation of the original galaxy into a single black hole may be slightly affected by gravitational radiation. When a number of stars go round each other, they radiate gravitational waves, thus lose energy and become more tightly bound. The time scale over which this process has a significant effect on the galaxy is anything from 10^{24} to 10^{30} years (Islam, 1977; Dyson, 1979). Thus the effects of dynamical evolution will be more dominant than those of gravitational radiation.

11.4 Black-hole evaporation

According to the laws of classical mechanics, a black hole will last forever. It was shown by Hawking (1975) that when quantum phenomena are taken into account, a black hole is not perfectly black but gives off radiation such as electromagnetic waves and neutrinos. 'Empty' space is actually full of 'virtual' particles and antiparticles that come into existence simultaneously at a point in space, travel a short distance and come together again, annihilating each other. The energy for their existence can be accounted for by the uncertainty principle. In the neighbourhood of the horizon of a black hole it might happen that one particle from a virtual pair falls into the black hole with negative energy, while its partner, unable to annihilate, escapes to infinity with positive energy. The negative energy of the falling particle causes a decrease in the mass of the black hole. In this manner the black hole gradually loses mass and becomes smaller, eventually to disappear altogether. The time scale for its disappearance is given by

$$t_{\mathrm{bh}} = G^2M^3/\hbar c^4. \tag{11.10}$$

For a black hole of one solar mass $t_{\mathrm{bh}} = 10^{65}$ years.

A black hole radiates as if it were a black body with a temperature which is inversely proportional to its mass. Such a black-body spectrum existed, as we have seen earlier, in the radiation in the early stages of the universe; it is describable in terms of a single temperature. The temperature of a black hole is of the order of $10^{26}/M\ K$ where M is the mass of the

black hole in grams. For a supergalactic black hole this amounts to about 10^{-18} K. If the temperature of the cosmic background radiation is higher than this, the black hole will absorb more energy than it radiates. But as the universe expands, the temperature of the background radiation, which is proportional to $(R(t))^{-1}$, decreases. In the Einstein–de Sitter universe a temperature of 10^{-20} K would be reached in 10^{40} years, whereas in the dust universe with $k=-1$ (where R is asymptotically proportional to t) this temperature would be reached in 10^{30} years. For models with a positive cosmological constant this temperature would be reached earlier, since for these models R behaves exponentially (asymptotically) with time. Thus by the time galactic and supergalactic black holes are formed, or some time afterwards, the temperature of the black holes will exceed that of the background radiation and they will begin to radiate more than they absorb.

From (11.10) we see that a galactic black hole will last for about 10^{90} years while a supergalactic black hole will evaporate completely in about 10^{100} years. Thus after 10^{100} years or so black holes of all sizes will have disappeared, that is, all galaxies as we know them today will have been completely dissolved and the universe will consist of stray neutron stars, black dwarfs and smaller planets and rocks that were ejected from the galaxies. There will be an ever-increasing amount of empty space in which there will be a minute amount of radiation with an ever-decreasing temperature.

11.5 Slow and subtle changes

Consider the long term behaviour of any piece of matter, such as a rock or a planet, after it has cooled to zero temperature. Its atoms are frozen into an apparently fixed arrangement by the forces of cohesion and chemical binding. But from time to time the atoms will move and rearrange themselves, crossing energy barriers by quantum mechanical tunnelling. Even the most rigid materials will change their shapes and chemical structure on a time scale of 10^{65} years or so, and behave like liquids, flowing into spherical shape under the influence of gravity.

Any piece of ordinary matter is radioactive because it can release energy by nuclear fusion or fission reactions which take place by quantum tunnelling. All pieces of matter other than neutron stars must decay ultimately to iron, which has the most stable nucleus. The life-time for decay is given approximately by the Gamow formula $\exp[Z(M/m)^{1/2}]$, where Z is the nuclear charge, M the nuclear mass and m the electron mass. To get the actual life-time one has to multiply this pure number by some typical

nuclear time scale, say 10^{-21} s. This gives a life-time of from 10^{500} to 10^{1500} years. On this time scale ordinary matter is radioactive and is constantly generating nuclear energy.

What will eventually happen to black dwarfs and neutron stars? If a black dwarf is compressed from outside by some external agent, it will collapse to a neutron star. In the near emptiness of the future universe there will be no external agent to compress it. However, the 'compression' can occur spontaneously by quantum tunnelling. The time scale can be calculated by another form of the Gamow formula, and is given as $10^{10^{76}}$ years (Dyson, 1979). In a similar period, a neutron star will collapse into a black hole by quantum tunnelling and eventually evaporate by the Hawking process. Thus ultimately all black dwarfs and neutron stars will also disappear.

The decay of black dwarfs and neutron stars (indeed, of smaller pieces of matter) may occur earlier than $10^{10^{76}}$ years if black holes of smaller than stellar size are possible. Let M_B be the minimum size of a black hole, that is, suppose it is not, in principle, possible for a black hole to exist with mass less than M_B. Then the following alternatives arise:

(*a*) $M_B = 0$. In this case all matter is unstable with a comparatively short life-time.

(*b*) M_B is equal to the Planck mass: $M_B = M_{PL} = (hc/G)^{1/2} = 2 \times 10^{-5}$ g. This value of M_B is suggested by Hawking's theory, according to which every black hole loses mass until it reaches a mass of M_{PL}, at which point it disappears in a burst of radiation. In this case the life-time for all matter with mass greater than M_{PL} is $10^{10^{26}}$ years, while smaller pieces are absolutely stable.

(*c*) M_B is equal to the quantum mass $M_B = M_Q = hc/Gm_P = 3 \times 10^{14}$ g, where m_p is the proton mass. M_Q is the mass of the smallest black hole for which a classical description is possible (Harrison, Thorne, Wakano and Wheeler, 1965). In this case the life-time for a mass greater than M_Q is $10^{10^{52}}$ years.

(*d*) M_B is the Chandrasekhar mass $M_{ch} \simeq 4 \times 10^{33}$ g. In this case the life-time for a mass greater than M_{ch} is $10^{10^{76}}$ years, as mentioned earlier.

The long term future of matter in the universe depends crucially on which alternative is correct. Dyson (1979) favours (*b*). In the analysis so far we are assuming that the 'stable' elementary particles such as electrons and protons are, in fact, stable. This may not be the case over the periods which we have been discussing.

Barrow and Tipler (1978) show, under certain assumptions, that the universe will become increasingly irregular and unstable against the development of vorticity. This conclusion, however, is based on the assumption that the universe will consist of pure radiation in the long run, with all matter decaying. The matter density of stable matter varies as R^{-3} while that of radiation varies as R^{-4}. Thus radiation will dominate only if all matter decays. It is not clear how far this assumption is justified. Page and McKee (1981) find that a substantial proportion of the electrons and positrons (the latter arising from the decay of protons) will never annihilate.

The concept of the passage of time loses some of its meaning when applied to the final stages of the universe. Time is measured against some constantly changing phenomena. The only way in which the passage of time will manifest itself finally will be, presumably, the density and temperature of the background radiation, which will approach zero but never quite reach it.

The long term future of life and civilization has been discussed by Dyson (1979) (see also Islam (1979a,b, 1983a), and Krauss and Starkman (1999)).

11.6 A collapsing universe

The long term future of the universe is very different if the universe stops expanding and starts to contract. The life-time for a closed universe depends on the present average density of the universe.

Suppose the present density of the universe is twice the critical density. The universe will then expand to about twice its present size and start to contract. The total duration of the universe will be about 10^{11} years. The cosmic background radiation will go down to about 1.4 K and start to rise thereafter. The turning point will come in a few tens of billions of years – there will not be much change in the universe during this time. After the turning point, all the major changes that took place in the universe since the big bang will be reversed. In a few tens of billions of years, the cosmic background temperature will rise to 300 K, and the sky will be as warm all the time as it is during the day at present. After a few million years, galaxies will mingle with each other and stars will begin to collide with each other at frequent intervals. But before they get disrupted by such collisions, they will, in fact, dissolve because of the intensity of the background radiation (Rees, 1969), which will eventually knock out all electrons from atoms and finally neutrons and protons from nuclei. Ultimately, there will be a universal collapse of all matter and radiation into a compact space of

infinite or near infinite density. It is not clear what will happen after such a collapse. Indeed, it is not clear if it is meaningful to talk about 'after' the final collapse, just as it is unclear whether it is meaningful to ask what happened 'before' the big bang.

In the steady state model proposed by Bondi and Gold (1948) and by Hoyle (1948) mentioned earlier, it is, in principle, possible for the universe to stay the same into the indefinite future. But as we have seen such a model is observationally untenable. It is also not clear in what way the above scenario is affected by the inflationary models, in which it appears possible to have different universes.

Appendix

A1. Introduction

In this appendix we consider topics some of which are extensions of material covered in the earlier chapters, and other additional ones which are not necessarily recent developments, but may have relevance for cosmological studies generally. We discuss both observational and theoretical matters.

A2. Neutrino types

A significant discrepancy between theory based on the standard model of particle physics and observation of the flux of solar neutrinos on the surface of the Earth has been noticed for some years. In spite of much effort, an adequate explanation of this discrepancy has not been found.

As discussed in Section 8.8, the number of types of neutrino is of cosmological importance. Among relevant points to emerge at the 14th International Conference on Neutrino Physics and Astrophysics at CERN in 1990 was that there are three neutrino types unless the mass of the fourth one exceeds 45 GeV; the relic abundance of such a heavy neutrino is not sufficient to contribute to dark matter (Griest and Silk, 1990; Salati, 1990). These results come from LEP, the Large Electron Positron collider at CERN.

A large detector has been set up at Mount Ikenoyama in an active zinc mine in Japan, known as the Super-Kamiokande Detector (Kearns, Kajita and Totsuka, 1999). The original experiment was concerned with the detection of proton decay, and was set up at Kamioka, a mining town about 250 km from Tokyo. The name 'Kamiokande' stands for 'Kamioka

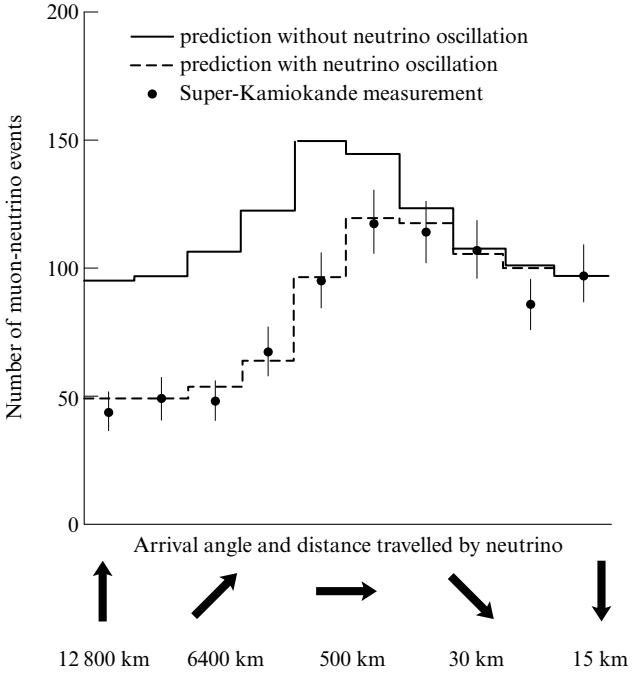


Fig. A1. This graph displays the number of high-energy muon-neutrinos arriving on different trajectories at Super-K, indicating neutrino oscillations. Upward-going neutrinos, plotted towards the left, have travelled far enough for half of them to change flavour and escape detection (after *Scientific American*, August 1999).

Nucleon Decay Experiment'. A similar experiment was the IMB one located in a salt mine near Cleveland, Ohio. Although no proton decays have been seen, the same experimental set-up is suitable for detecting neutrino oscillations, because hundreds of events have been recorded of neutrino interactions. Super-Kamiokande, or Super-K, is a similar machine but about ten times bigger. Interesting data are beginning to emerge, with some corroboration from experiments carried out elsewhere. These indicate that muon-neutrinos transform into other kinds, perhaps tau-neutrinos. The expected flux of muon-neutrinos, which include those coming through the Earth from below as well as those coming from above, which should be about twice that of the electron-neutrino flux, amounts to only 1.3 times instead.

Figure A1 gives a graph indicating this discrepancy. Neutrino oscillation, as stated earlier, indicates mass; the present discrepancy leads to mass

of the heavier neutrino of 0.03 to 0.1 eV. This is small enough to be accommodated in the Standard Model.

An experiment carried out at Los Alamos National Laboratory detected electron-neutrinos from a source that is meant to produce only muon-neutrinos, indicating oscillations. These results, interesting as they are, will be clarified by further experiments at these laboratories, and other ones such as the Sudbury Neutrino Observatory in Ontario and the Chooz nuclear power station in Ardennes, France. Theoretically also there are various alternative possibilities which have to be carefully examined. These matters are relevant to aspects of cosmology as well as to particle physics.

A3. A critique of the standard model

Arp, Burbidge, Hoyle, Narlikar and Wickramasinghe (1990) are very critical of the standard model as described in the previous chapters and as believed by a great majority of cosmologists. Arp *et al*. cite various pieces of evidence to support their contention that, 'perhaps, there never was a Big Bang'. They also claim that the large red-shifts discovered so far, or at least substantial portions thereof, are in fact a result of intrinsic properties of the sources so that they do not lie at large cosmological distances, but are much closer, at distances that would follow from the Hubble Law for red-shifts $z \leqslant 0.1$. One of the reasons for this view is the discovery by Arp *et al*. of cases of galaxies of very different red-shifts which are found very close together on the photographic plate. Opponents of this view contend that these are purely chance alignments of galaxies which are in reality very far from each other. Arp *et al*. are aware of this criticism but they insist that their findings are statistically significant. Arp *et al*. discuss at length the various other reasons for their lack of belief in the standard model. For example, they claim that the cosmic background radiation is not a relic of the primordial big bang, but is a result of the thermalization (that is, attainment of black-body spectrum) of the radiation given off after galaxy formation, and they suggest mechanisms through which thermalization could have occurred. They admit that they have no clear alternative for the standard model, but they suggest that a variation of the steady state model (see Section 8.3), which can be considered as one of the forms of the scale-invariant conformal theory of gravitation put forward by Hoyle and Narlikar (see, e.g., Hoyle and Narlikar, 1974), fits the current observations, as interpreted by them, better. Various points Arp *et al.* discuss are of intrinsic interest, whether or not their overall view is correct. Although this is a minority and an unpopular view, we believe such criticism is healthy for

the subject of cosmology, for no theory or model should turn into a set of dogmas (Oldershaw, 1990). The onus is on the adherents of the standard model to provide adequate answers to these criticisms. Presumably some adherents would claim that adequate answers have already been given, but one can expect more answers to appear in the near future.

Hoyle, Burbidge and Narlikar have explained in detail this critique of the standard model in an interesting book (Hoyle, Burbidge and Narlikar, 2000), which is of considerable importance to cosmologists, even if they don't agree with the critical point of view. Chapters such as those entitled 'The observational trail 1931–56, the determination of H_0 and the age dilemma', 'The extension of the redshift-apparent magnitude diagram to faint galaxies 1956–95', 'The cosmic microwave background – an historical account', 'The origin of the light elements', and others, by three experienced cosmologists, are extremely valuable for students of cosmology of all opinions. One can hope that the publication of this book will stimulate critical examination of various aspects of cosmology and lead to genuine progress.

A4. An accelerating universe?

Since the last two years or so evidence appears to be accumulating for the existence of a positive cosmological constant, which would imply an accelerating universe. There is some support for the latter possibility from a detailed study of the spectrum of the cosmic background radiation (Perlmutter *et al*., 1998; Krauss, 1998, 1999). This circumstance, which some regard as a revolution in observational cosmology, has arisen mainly due to major improvement in techniques for observing supernovae explosions in distant galaxies, which had not been hitherto possible. The method involves surveying the sky with powerful optical telescopes at intervals of a few days and making a detailed comparison to see if any galaxies display brightening. In this manner it is possible to detect numerous Type Ia (SNe Ia) supernovae, whose absolute luminosities are known to within a reasonable range. The red-shift can be measured and so an analysis can be carried out which can provide information about the evolution of the universe in earlier epochs, some billions of years ago. This programme has, of course, been carried out for decades, but never before has anything like the present accuracy been attained in the measurement of light which left the objects concerned at an earlier time which is a significant fraction of the age of the universe. (See also Branch, 1998; Hogan, Kirshner and Suntzeff, 1999.)

If the new observations are valid and confirmed, the implications for cosmology, needless to say, are very important. The observations will doubtless be repeated many times in the next few years, and the results of these observations will be eagerly awaited by all cosmologists. At the same time, on the theoretical front, the causes for a positive cosmological constant, if indeed there is one, will have to be assiduously searched. Various reasons have been given, such as vacuum fluctuations (by Zel'dovich; see Weinberg, 1989), but these arguments are tentative and there are difficulties with each. Doubtless Einstein would have been intrigued by these developments!

A5. Particle physics and quantum field theory

In the last few years, an intimate connection has developed between cosmological studies and the theory of elementary particles, especially with regard to the early, very early universe and the origin of the universe. A relatively non-technical account of this connection has been given in the chapters on the early and very early universe. A somewhat more technical account of an aspect of this connection has been presented in the chapter on quantum cosmology. There are many good books containing the technical material required on various aspects of quantum field theory – quantum electrodynamics, and gauge theories such as the Glashow–Weinberg–Salam electro-weak theory and quantum chromodynamics. The older approach of canonical quantization is described in standard books by Schweber (1961) and by Bjorken and Drell (1965) among others, while path integral quantization, more suitable for gauge theories, is discussed in books by Ryder (1996) and Itzykson and Zuber (1980). The preliminary account of path integrals and of the Schrödinger functional equation given here may be useful in this context, in a small measure. Relatively non-technical but useful accounts of these and related topics are contained in reviews by Salam (1989), Taylor (1989) and others.

In this section we shall describe briefly an important ingredient that forms a part of these considerations, namely, Feynman diagrams; these have been mentioned in Chapter 8. Feynman diagrams can be derived either from the canonical quantization of fields, or from the path integral formalism. We mention the result, as described on pp. 229–232 of the book by Ryder (1996). The two second-order diagrams (with two vertices) displayed in Fig. A2 contribute to pion-nucleon scattering (these are to be read from left to right, unlike the diagrams of Fig. 8.2 which go from the bottom to the top). These involve interaction of a (pseudo-)scalar particle (pion) with a spinor particle (nucleon). For example, in $\pi^+ p$ scattering, the

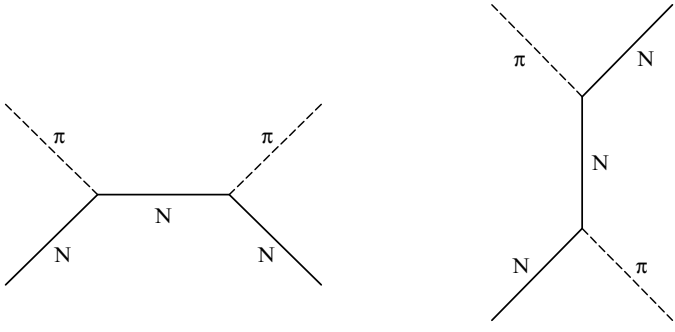


Fig. A2. These two second-order diagrams contribute to the amplitude for pion-nucleon scattering. They are to be read from left to right, unlike the diagrams of Fig. 8.2, which go from the bottom to the top. Here, N represents a nucleon.

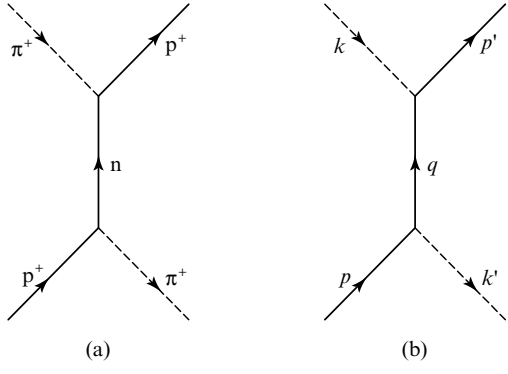


Fig. A3. These are two versions of the diagram of Fig. A2(b) for the case of π^+p scattering; (a) displays the particles and (b) represents the four-momenta of the same particles. In (a) the charge on the proton is indicated to distinguish it from the four-momentum p; n is the neutron.

diagram of Fig. A2(b) could be represented by those of Fig. A3; Fig. A3(a) displays the particles and Fig. A3(b) the four-momenta of the same particles. For a scalar or a pseudoscalar particle interacting with a spinor particle, the following rules apply for constructing the nth-order Feynman diagram (with n vertices) (the spinors $u(p)$, etc., are defined below):

(1) The amplitude for a particular process, with specified ingoing and outgoing particles, for a particular order is obtained by adding the amplitudes for all topologically inequivalent connected diagrams;

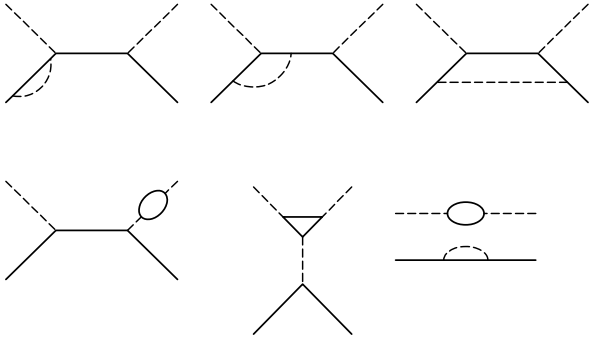


Fig. A4. Some fourth-order diagrams for scalar/pseudoscalar-spinor (e.g. pion-nucleon) scattering. The last diagram is disconnected and is not counted (from Ryder, 1996, p. 231).

Fig. A2 gives two for the second order, Fig. A4 displays some fourth-order diagrams. Scalar lines are dotted, spinor lines continuous.

(2) Each incoming spinor particle corresponds to a factor $u(p)$ ($v(p)$ for its antiparticle), and each outgoing spinor particle to a factor $\bar{u}(p)$.

(3) With each vertex goes a factor ig (for scalar interaction) or $ig\gamma_5$ (for pseudescalar), with g as the suitable coupling constant (occurring in the Lagrangian); and a factor $(2\pi)^4\delta^4$ (incoming momenta).

(4) For each internal continuous line, i.e., a spinor propagator of momentum p, insert the factor:

$$\frac{1}{(2\pi)^4}\frac{i}{(\gamma p - M)}\,\mathrm{d}^4p;\ \gamma p = \gamma^\mu p_\mu. \tag{A1}$$

(5) For each (pseudo-)scalar propagator, include the factor:

$$\frac{1}{(2\pi)^4}\frac{i}{(p^2 - m^2)}\,\mathrm{d}^4p. \tag{A2}$$

(6) Integrate over internal momenta.

Following these rules, the contribution to π^+p scattering from the diagram of Fig. A3 can be written as follows (this is Ryder's Eq. (6.173)) (the indices s', s on $\bar{u}$ and u take values 1, 2 and refer to different spinors; see Ryder's Eq. (2.139)):

$$s_{\mathrm{fi}} = -2\mathrm{i}\delta^4(P_\mathrm{f} - P_\mathrm{i})g^2(2\pi)^4\bar{u}^{s'}(p')\,\gamma_5\,\frac{\gamma.(p-k')+M}{(p-k')^2 - M^2}\,\gamma_5 u^s(p), \tag{A3}$$

where P_i, P_f are respectively the total initial and final four-momentum; the delta-function implies conservation: $P_i = p + k = P_f = p' + k'$. The spinors $u(p)$, etc., are Fourier transforms of the spinors ψ, in configuration space, which are solutions of the Dirac equation:

$$(i\gamma^\mu \partial_\mu - m)\psi(x) = 0, \tag{A4}$$

γ^μ being the Dirac matrices (γ^0, γ^1, γ^2, γ^3), with $\gamma_5 = i\gamma^0\gamma^1\gamma^2\gamma^3$. The $\bar{u}$, $\bar{\psi}$ are conjugate spinors defined in terms of the complex conjugate of the components of u, ψ. The details of these functions, and the manner in which actual cross sections can be derived from the functions represented by the diagrams, can be found in the lucid book by Ryder (1996), which explains many aspects of particle physics and quantum field theory.

We indicate briefly how Feynman diagrams can be derived from the path integral formalism, for self-interacting scalar fields. The path integral over coordinates defined by (10.73) and over the metric as in (10.76) can be generalized to an integral (more appropriately called a 'functional integral') over, say, a scalar field interacting with itself through a Lagrangian such as that used in (10.17) (or (10.24)), as follows:

$$W[J] = \int \mathcal{D}\phi \exp\{i \int d^4x [\mathcal{L} + J(x)\phi(x)]\}, \tag{A5}$$

where a 'source' term $J(x)\phi(x)$ has been added to $\mathcal{L}$ so that the functional integral becomes a functional of $J(x)$. If one now takes repeated functional derivatives of $W[J]$ with respect to $J(x_1)$, $J(x_2)$, . . . for different space-time points x_1, x_2, . . . and sets $J = 0$, one obtains the usual Green's functions of quantum field theory. The Fourier transforms of these functions then yield the familiar Feynman diagrams when expanded in a suitable power series. Some problems arise about making the integral (A5) well defined. These may be dealt with by going over to Euclidean space with imaginary time: (τ, x), $\tau = ix_0$, or by adding to the integrand in the exponent an imaginary term quadratic in $\phi(x)$: $[\mathcal{L} + (1/2)i\epsilon\phi^2 + J(x)\phi(x)]$ with ϵ a small positive constant.

A6. Cosmic background radiation

One of the important observations relevant to cosmology was that carried out by the Cosmic Background Explorer (COBE) satellite (Lindley, 1990a; see also Carr, 1988; Hogan, 1990). This satellite carried an instrument which was especially designed to measure the departure in the cosmic background radiation from a smooth 'reference' black body. As indicated earlier, any deviation from a smooth background, that is, any 'graininess'

that is found, and its magnitude, can give useful information about primordial galaxy formation or other similar characteristics of the early universe. The range of wavelengths over which measurements were taken by the satellite was from 100 μm to 1 cm. It was found that departures from a black-body spectrum, if any, are less than 1%. The observations by COBE, the results of which were presented at the April 1992 meeting of the American Physical Society in Washington DC (see *News and Views*, *Nature*, *Lond.* **356**, 741 (30 April 1992)), reveal slight departures from uniformity, the variation in temperature ΔT being given by $\Delta T/T = (5 \pm 1.5) \times 10^{-6}$, over angles up to 90°. This is an extremely important observation which is likely to have a significant effect on theories of galaxy formation.

A7. Quasar astronomy

A significant advance in quasar astronomy (see Section 5.3) has been the observation of the optical spectra of the quasar Q1158+4635 (red-shift $z = 4.73$) and ten other quasars, with red-shifts $z > 4$ carried out by Schneider, Schmidt and Gunn (1989). Detailed statistical analysis remains to be done; these analyses are likely to provide clues to the physical conditions obtaining in the intergalactic medium in the very early evolution of the universe. An analysis of the fine structure in the absorption spectrum of a strong distant source such as a quasar can give useful information on types and concentrations of the intervening mass. This could possibly provide some clue to the problem of 'missing' or 'dark' matter.

A8. Galactic distribution

Broadhurst *et al.* (1990) (see also Davis, 1990) have studied large-scale distribution of galaxies at the galactic poles, both north and south. They find indications that galaxies are not distributed randomly but are clustered on scales of $5h^{-1}$ Mpc, where h is a constant denoting the uncertainty in the value of H_0; $H_0 = 100h$ km s^{-1} Mpc^{-1}, with a likely value in $0.5 \leqslant h \leqslant 1$. For this survey, data are taken from four different surveys at the north and south galactic poles. They find indications of periodic oscillations of density and evidence of structure at the largest scale studied by them. They emphasize the tentative nature of these observations, which need to be confirmed. If confirmed, these observations may have implications for theories of galaxy formation and for inflationary models.

A9. Value of H_0 using planetary nebulae

There has been an estimate of the value of H_0 by Jacoby, Ciardullo and Ford (1990) (see also Fukugita and Hogan, 1990) which seems to be of considerable interest. As is clear from the earlier chapters, a correct observational determination of the value of H_0 is one of the most important problems in cosmology. As indicated earlier, the main difficulty here is to determine accurately the distance of galaxies which are relatively far; this is usually done by comparing their luminosity with that of standard candles such as Cepheid variables and Type Ia supernovae. The former exist only for nearby galaxies, while the latter are rare events. Jacoby *et al.* have been able to determine the distance to several galaxies in the Virgo cluster more accurately than before with the use of another type of standard candle, namely, planetary nebulae. The latter are clouds of radiating gas to which a star usually transforms towards the end of its life, when its hydrogen fuel is exhausted and it is burning only helium. The interesting thing is that there seems to be a maximum intrinsic brightness associated with planetary nebulae, the theoretical reason for which is not entirely clear; this could be to do with the maximum mass of the core of a star nearing its end – one which does not become a neutron star or a black hole – the so-called Chandrasekhar mass (around 1.4 solar masses) (see Section 11.2). Another advantage of the technique used by Jacoby *et al*. seems to be that planetary nebulae seem to emit most of their energy in a narrow spectral band. This results in ease of detection and necessity of observing at a single epoch, unlike Cepheids. Hitherto the value of H_0 has been uncertain by a factor of about 2. Jacoby *et al.* claim to have calculated H_0 to within 15% in the range 75–100 km s^{-1} Mpc^{-1}, which is in the higher range of the previous uncertainty of 50–100 km s^{-1} Mpc^{-1}. This would have serious implications for cosmology, if confirmed. For example, this would imply that the universe is somewhat younger than previously believed. (See (4.4), (4.25) and Section 4.2).

A10. Cosmic book of phenomena

Peebles and Silk (1990) have compiled an interesting 'Cosmic book of phenomena' comparing five general theories for the origin of galaxies and large-scale structure in the universe by studying how well these theories are able to explain 38 different observational phenomena. This follows their earlier 'book' (1988) which dealt exclusively with large-scale structure. As

mentioned in Chapter 4, estimates of the value of Ω, the density parameter (see below (4.9)) based on observations and on the dynamics of systems of galaxies, yield a value somewhat less than unity, around 0.1. Theorists prefer a value close to unity, for reasons given in Section 9.1 ((9.1a), (9.1b)). The two points of view here are therefore, roughly speaking, firstly, that $\Omega \approx 0.1$ with the mass density consisting mainly of ordinary (baryonic) matter and, secondly, that the universe is dominated by some exotic non-baryonic matter which interacts weakly so that it is not readily detected (dark matter). Peebles and Silk examine the following five general theories which purport to explain the above scenarios, by seeing how well they deal with 38 different observational constraints. The first is the cold dark matter (CDM) theory (Frenk *et al,* 1988, 1990) in which the universe is Einstein–de Sitter (see Section 4.2), dominated by matter with negligible initial pressure (cold matter) that interacts weakly, and galactic structure emerges through suitable primeval density fluctuations. The hot dark matter (HDM) model (Zel'dovich, Einasto and Shandarin, 1982) has particles of dark matter with primeval velocity typical of neutrinos of mass about 30 eV; the remnant neutrinos make $\Omega \approx 1$ (see Section 8.8). In the string theories (STR) structure is formed by seeds of primeval non-linear perturbations; we shall come back to these theories. Weinberg, Ostriker and Dekel (1989) attempt to explain the origin of structure in what Peebles and Silk call the explosion (XPL) picture, in which locally inserted energy, which could be from early supernovae, creates ridges of baryons which subsequently disintegrate to form new star clusters. In the baryonic dark matter (BDM) theory, unlike in the CDM theory, most of the galaxy masses were assembled at red-shifts $z \geqslant 10$. Peebles and Silk define a 'quality rating' parameter r, as follows

$$r = \tfrac{1}{2}(1 + 2wp - w), \tag{A.6}$$

where p is the probability for the theory and w is the weight for the phenomenon that is being explained. The parameter r has the character of a probability. If the weight w for the phenomenon is high, $w \approx 1$, then the rating r is nearly the same as p, the probability that the theory explains the phenomenon. If the weight is very low, $w \approx 0$, then $r \approx 0.5$, independent of p. Other cases fall in between these extreme cases. Peebles and Silk combine the r_i for 38 phenomena and compute the product Πr_i, which is then used to determine the overall rating. An improbable theory would have a significant number of small r_i, whereas a 'good' theory would have more r_i near unity. Peebles and Silk find no clear winners but the CDM and BDM theories seem to them to be slightly ahead of the rest. As exam-

ples, we consider two of the phenomena in the list and give the weights w and the ratings r. The first one is that the isotropy of the cosmic background radiation is given by $\Delta T/T < 2 \times 10^{-5}$ at around 30 arcmin. The weight w and the rating r for the five theories CDM, HDM, STR, XPL and BDM are respectively 1.0, 0.95, 0.05, 0.70, 0.70, 0.70. The second one is that, for the phenomenon that there are clusters of galaxies as massive as the Coma cluster at $z = 1$, these quantities have the values 0.8, 0.14, 0.42, 0.86, 0.86, 0.86, respectively. These two phenomena are taken at random from the list; there are 38 such phenomena in the list, as mentioned earlier.

A11. Cosmic strings

As mentioned earlier (see Section 9.5), among the possible relics of the phase transition of the very early universe are cosmic strings, which can be considered as thin lines of concentrated energy. If cosmic strings exist, they could be important for the formation of galaxies and large-scale structure of the universe. The evidence for cosmic strings is hard to find; this could come, for example, from gravitational radiation, which is notoriously difficult to detect. To be important for galaxy formation the mass per unit length of the strings should be in the region of 10^{22} g cm^{-1}, which is roughly the magnitude predicted by GUT. Such densities would produce certain potentially detectable observational effects, such as double images of distant galaxies and quasars due to gravitational lensing, certain discontinuities in the microwave background radiation, in addition to effects on gravitational radiation mentioned. The theoretical discussion of cosmic strings is difficult and interesting; they form a tangled web permeating the entire universe, with closed loops or extending to infinity without ends. Their evolution is believed to be scale-invariant; statistically the network is the same at all times. This implies that at any time t, the distance between nearby long strings is of the order of the horizon $\sim ct$ and typical loop size is a certain fraction of this distance. Simulations (Bennett and Bouchet, 1989; Allen and Shellard, 1990; see also Vilenkin, 1990) show that long strings have a significant fine structure on a scale somewhat smaller than the horizon, contrary to what was believed earlier. A major portion of the structure is in the sharp angles, 'kinks', at points where strings are reconnected. The typical loop size l is also smaller than expected: $l \ll ct$. The new findings have interesting consequences for galaxy formation. There are two ways cosmic strings are believed to assist galaxy formation: gravitational attraction of loops, and formation of wakes behind fast-moving long strings; these were thought to be comparable.

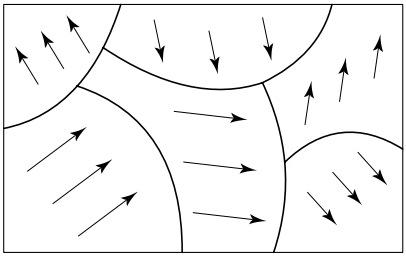


Fig. A5. When a ferromagnet is cooled below 1043 K, domains form with different magnetization.

The new studies indicate, since loop sizes may be much smaller, that the second process may dominate. Further studies are needed to clarify various aspects of this interesting point.

A12. Topological structures

Turok (1989) (see also Friedman and Morris, 1990) has considered topological structures in the very early universe. As has been noted earlier, the astonishing uniformity of the cosmic background radiation is difficult to reconcile with the clumping of matter into galaxies and clusters. In Chapter 8 we saw that the background radiation is composed of radiation that left matter about 100 000 years after the big bang. As this radiation is isotropic to 1 part in 10^4 or so, the density variation around the period the radiation left matter could not have been significantly more than this fraction. It is difficult to evolve galaxies with such small variations unless one has exotic forms of matter. (In fact Arp *et al.*, 1990, quoted earlier, cite this as a reason why galaxies should have been formed before the background radiation, although it is not clear if they can explain the extraordinary smoothness of the radiation.) Turok (1989) suggests that topological structures related to strings could provide ingredients for the formation of galaxies. It was mentioned in Chapter 9 that the symmetry of the four forces, namely, gravitation, electromagnetic, weak and the strong forces, presumably was broken successively through phase transitions in the very early universe. In addition to the example of freezing water cited in Chapter 9, one can consider the breaking of symmetry when a ferromagnet is cooled below 1043 K; this results in alignment of the randomly oriented spins, which form distinct domains, as in Fig. A5. As mentioned in Chapter 9, a similar breaking of symmetry may have occurred in the very

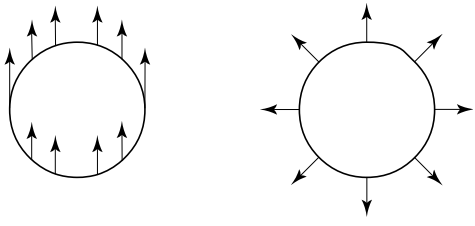


Fig. A6. The two configurations (*a*) and (*b*) cannot be deformed into one another through continuous transformations.

early universe, which may be considered as being due to the appearance of Higgs fields. A topological structure may be associated with a Higgs field, which can be understood by considering a vector field pervading the universe, represented by an arrow of unit length at every point of the universe. Two configurations have different topologies if they cannot be deformed into each other by continuous changes. For example, if we consider a one-dimensional 'universe' (e.g., a circle), then the two configurations (*a*) and (*b*) of Fig. A6 cannot be continuously deformed into each other. These two configurations have different 'winding numbers'. (A typical element U of a groups such as $SU(2)$ can be regarded as a mapping from S^3, the three-dimensional surface of a sphere in four-dimensional Euclidean space, onto the group manifold of $SU(2)$, that is, the space of parameters characterizing the group $SU(2)$, which is topologically the same as S^3. The *winding number* of a particular class of mappings is the number of times the spatial S^3 is covered by the group manifold S^3. Gauge transformations belonging to a group G, which can be any $SU(N)$, can be split up into homotopy classes, each of which is characterized by a distinct winding number). Essentially what may happen is that as a universe with a certain topological structure evolves, because this structure is preserved (cannot be made to 'go away') one may eventually get small regions of high energy density (called 'knots'), which may provide seeds for galaxy formation.

A13. Extended inflation

In Chapter 9 we saw that one of the properties of the observed universe the inflationary models attempt to explain is the fact that Ω, the density parameter (see (9.1a), (9.1b)) is so close to unity. As mentioned earlier, the essential idea is that the universe spends a very short period in its

very early history in a supercooled state, when a large constant and positive vacuum energy dominates its density of energy. The subsequent exponential expansion causes Ω to evolve towards unity. Also, inflation expands a causally connected region that is small into one that is much larger than the observable universe, thus solving the 'horizon' problem. In the 'old inflation' of Guth, there were 'bubbles' of the true vacuum in the supercooled state which could not merge and complete the phase transition. In the 'new inflation' this problem could perhaps be solved, but this required such 'fine tuning' of the parameters that it was not clear that such fine tuning could be achieved. Steinhardt (1990), proposes a model that he calls 'extended inflation' (see also Lindley, 1990b), which, it is claimed, does not have the defects of earlier models in that there exist ranges of parameters which allow a set of initial conditions that lead to $\Omega \leqslant 0.5$, so that consistency with observation is obtained. As in 'old inflation', in 'extended inflation' the barrier between the false and true vacuum is finite, but the new feature here is that the strength of gravitation varies with time, and this variation is related in a certain sense to the expansion of the universe. Steinhardt also shows that in the earlier 'new inflation' the fine tuning looked for could not have been achieved.

A14. Quantum cosmology

In Chapter 10 on quantum cosmology it was stated that the expression (10.37) for the amplitude has hidden in it many complexities, one of these being similar to that encountered in Yang–Mills theories which was dealt with by Faddeev and Popov (1967). In fact, because of the indefinite metric and the nature of the space of geometries over which the path integral is taken, other complications arise of a different nature from that encountered in Yang–Mills theories. A satisfactory and precise formulation and definition of (10.37) (see also (10.55), (10.56)) still remains an important problem in quantum cosmology (see Halliwell and Hartle, 1990; Halliwell and Louko, 1989a,b). An important aspect of the problem of quantum cosmology is that of 'decoherence', that is, the nature of the interference between different histories of the universe and the manner in which these effects eventually disappear to leave the universe to evolve classically subsequently (Gell-Mann and Hartle, 1990; see also Calzetta and Mazzitelli, 1990).

We make some additional remarks about the Wheeler–De Witt equation; some of the earlier steps may be repeated, for convenience. As before,

we set $\hbar = c = 1$ and introduce coordinates so that the space-like hypersurfaces are $t =$ constant and the metric is written as follows (10.63):

$$ds^2 = (N^2 - N_i N^i)dt^2 - 2N_i dx^i dt - h_{ij} dx^i dx^j, \qquad i, j = 1, 2, 3. \quad \text{(A7)}$$

The three-vector N^i is a contravaniant three-vector with respect to purely spatial transformations of (x^1, x^2, x^3) and N_i is the corresponding covariant three-vector derived with the use of the three-metric h_{ij}; N is a function defined below. Again, K_{ij} is the extrinsic curvature of the three-surface $t =$ constant, given by (10.64), where n^μ is the unit normal to the hypersurface $t =$ constant, n_j being the spatial part of the covariant components of this vector. The quantities K_{ij} can be evaluated in terms of N, N_i and h_{ij} as follows (see, e.g., Misner, Thorne and Wheeler, 1973, p. 513). Note first that the contravariant components of the metric $g^{\mu\nu}$ corresponding to $g_{\mu\nu}$ given by (A7) can be written as follows (we first write $g_{\mu\nu}$):

$$g_{00} = N^2 - N_i N^i, \; g_{0i} = -N_i, \; g_{ij} = -h_{ij}, \quad \text{(A8a)}$$

$$g^{00} = 1/N^2, \; g^{0i} = -N^i/N^2, \; g^{ij} = (-h^{ij} + N^i N^j/N^2), \quad \text{(A8b)}$$

where h^{ij} is the inverse of h_{ij} and, as mentioned, N_i, N^i are related through h_{ij}, that is,

$$h^{ik} h_{kj} = \delta^i_j, \; N_i = h_{ij} N^j. \quad \text{(A8c)}$$

We leave it as an exercise for the reader to verify, that the $g^{\mu\nu}$ given by (A8b) is the inverse of (A8a).

Next we show that the unit normal n^μ can be taken as follows:

$$n^\mu = (1/N, -N^i/N), \quad \text{(A9a)}$$

with

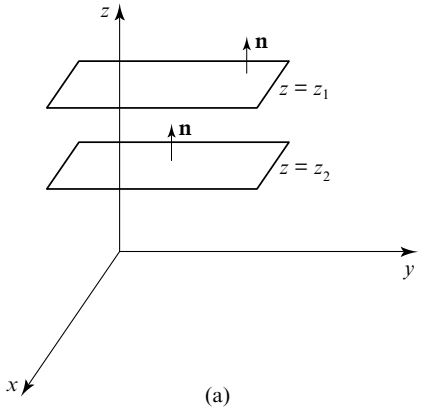
$$n_\mu = (N, 0, 0, 0), \quad \text{(A9b)}$$

as can be verified with the use of (A8a–c). A vector within the surface $t =$ constant can be taken as $m^\mu = (0, \Delta x^1, \Delta x^2, \Delta x^3) \equiv (0, \Delta x^i)$. It is then readily verified, with the use of (A8a–c) and (A9a,b), that

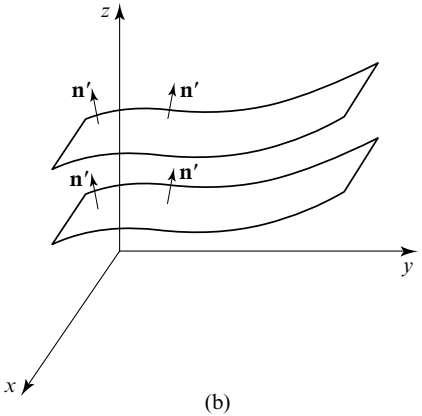
$$g_{\mu\nu} n^\mu n^\nu = 1, \; g_{\mu\nu} n^\mu m^\nu = 0, \quad \text{(A10)}$$

verifying both that n^μ is normal to the surface $t =$ constant and that it is of unit length. To go back to (10.64), we define the second-rank tensor $K_{\alpha\beta}$ as follows:

$$K_{\alpha\beta} = -n_{\beta;\alpha}. \quad \text{(A11)}$$



(a)



(b)

Fig. A7. Two-dimensional slices of three-dimensional Euclidean space with (a) constant normal (b) variable normal.

With the use of (A8a,b), (A11) and (A9a,b), we find

$$
\begin{aligned}
K_{ij} &= -n_{j;i} = -n_{j,i} + n_{\lambda}\Gamma^{\lambda}_{ji} \\
&= n_0\Gamma^0{}_{ji} = \tfrac{1}{2}Ng^{0\mu}(g_{\mu j,i} + g_{\mu i,j} - g_{ij,\mu}) \\
&= \tfrac{1}{2}Ng^{00}(g_{0j,i} + g_{0i,j} - g_{ij,0}) \\
&+ \tfrac{1}{2}Ng^{0k}(g_{kj,i} + g_{ki,j} - g_{ij,k}) \\
&= \frac{1}{2N}\left(-N_{j,i} - N_{i,j} + \frac{\partial h_{ij}}{\partial t}\right) \\
&+ \frac{N^k}{2N}(h_{kj,i} + h_{ki,j} - h_{ij,k})
\end{aligned}
$$

$$= \frac{1}{2N}(-N_{j,i} - N_{i,j} + \dot{h}_{ij}) + \frac{N_l}{N}\Gamma'^{\ell}_{ij}$$

$$= \frac{1}{2N}(-N_{j|i} - N_{i|j} + \dot{h}_{ij}). \tag{A12}$$

Here Γ'^{ℓ}_{ij} denotes the Christoffel symbol derived from the metric h_{ij} and a vertical stroke denotes covariant differentiation defined with the use of Γ'^{l}_{ij}, denoted by ∇_j in Section 10.7

The three-metric h_{ij} incorporates the intrinsic geometry of the surfaces $t =$ constant, while the extrinsic curvature K_{ij} determines how these surfaces are embedded in the four-dimensional space-time manifold. A simple example may help to clarify this situation. The ordinary three-dimensional Euclidean space may be 'sliced' into two-dimensional sections by the planes $z =$ constant (see Fig. A7(a), for which the unit normal $\mathbf{n}$ is constant, being the vector $\mathbf{k} = (0, 0, 1)$). Different 'slicings' are, however, possible, such as the one indicated in Fig. A7(b), where the intrinsic geometry of the two-dimensional sections remains the same as that of the plane, but the normal $\mathbf{n}'$ is now a function of position. The extrinsic geometry (determined by quantities corresponding to K_{ij}) is different in the two cases, and determines the manner in which the sections are embedded in the three-dimensional space. However, for the spatially closed universes considered here, these considerations do not apply directly, for it is difficult to define an intrinsic measure that locates the space-like hypersurface, apart from its intrinsic or extrinsic geometry (Hartle and Hawking, 1983).

Various aspects of quantum cosmology are described in an interesting book by D'Eath (1996).

Bibliography

Adams, E. N. 1988, *Phys. Rev.* **D37**, 2047.
Adams, F. C., Freese, K. and Widrow, L. M. 1990, *Phys. Rev.* **D41**, 347.
Albrecht, A. and Brandenberger, R. H. 1985, *Phys. Rev*. **D31**, 1225.
Albrecht, A. and Steinhardt, P. 1982, *Phys. Rev. Lett.* **48**, 1220.
Allen, B. and Shellard, E. P. S. 1990, *Phys. Rev. Lett.* **64**, 119.
Arp, H. C. 1967, *Ap. J.* **148**, 321.
Arp, H. C. 1980, in *Ninth Texas Symposium on Relativistic Astrophysics*, eds. J. Ehlers, J. Perry and M. Walker: New York Academy of Sciences, p. 94.
Arp, H. C., Burbidge, G., Hoyle, F., Narlikar, J. V. and Wickramasinghe, N. C. 1990, *Nature* **346**, 807.
Azad, A. K. and Islam, J. N. 2001, 'An Exact Inflationary Solution for an Eight Degree Potential', preprint.
Baade, W. 1952, *Trans. IAU* **8**, 397.
Bagla, J. S., Padmanabhan, T. and Narlikar, J. V. 1996, *Com. Astrophys.* **18**, 275.
Bahcall, J. N. and Haxton, W. C. 1989, *Phys. Rev.* **D40**, 931.
Bahcall, J., Huebner, W., Magee, N., Mertz, A. and Ulrich, R. 1973, *Ap. J.* **184**, 1.
Bahcall, J. *et al*. 1980, *Phys. Rev. Lett.* **45**, 945.
Banks, T. 1988, *Nucl. Phys.* **B309**, 493.
Barrow, J. D. 1987, *Phys. Lett*. **B187**, 12.
Barrow, J. D. 1990, *Phys. Lett.* **235**, 40.
Barrow, J. D. 1993, *Times Higher Education Supplement*, May 14.
Barrow, J. D. and Liddle, A. R. 1997, *Gen. Rel. Grav.* **29**, 1503.
Barrow, J. D. and Maeda, K. 1990, *Nucl. Phys.* **B341**, 294.
Barrow, J. D. and Saich, P. 1990, *Phys. Lett.* **249**, 406.
Barrow, J. D. and Tipler, F. J. 1978, *Nature* **276**, 453.
Baum, W. A. 1957, *Ap. J.* **62**, 6.
Beer, R. and Taylor, F. W. 1973, *Ap. J.* **179**, 309.
Belinskii, V. A. and Khalatnikov, I. M. 1969, *Soviet Phys. JEPT* **29**, 911.
Belinskii, V. A., Khalatnikov, I. M. and Lifshitz, E. M. 1970, *Adv. Phys*. **19**, 525; 1970, *Soviet Phys. JEPT* **33**, 1061.
Bennett, D. P. and Bouchet, F. R. 1989, *Phys. Rev. Lett*. **63**, 2776.
Birkinshaw, M. and Hughes, J. P. 1994, *Ap. J.* **420**, 33.
Bjorken, J. D. and Drell, S. D. 1965, *Relativistic Quantum Fields*, McGraw-Hill.
Black, D. C. 1971, *Nature Phys. Sci.* **234**, 148.
Black, D. C. 1972, *Geochim. Cosmochim. Acta* **36**, 347.
Boato, G. 1954, *Geochim. Cosmochim. Acta* **6**, 209.

Bogoliubov, N. N. and Shirkov, D. V. 1983, *Quantum Fields*, Benjamin-Cummings Publishing Company, Inc., Reading, Massachusetts.
Bolte, M. 1994, *Ap. J.* **431**, 223.
Bolte, M. and Hogan, C. J. 1995, *Nature*, **371**, 399.
Bond, J. R. 1992, in *Highlights in Astronomy*, Vol. 9, Proceedings of the IAU Joint Discussion, ed. J. Berjeron, Kluwer, Dordrecht.
Bond, J. R. 1995, *Phys. Rev. Lett.* **74**, 4369.
Bond, J. R., Cole, S., Efstathiou, G. and Kaiser, N. 1991, *Ap. J.* **379**, 440.
Bondi, H. 1961, *Cosmology*, Cambridge University Press, Cambridge, England.
Bondi, H. and Gold, T. 1948, *Mon. Not. Roy. Astr. Soc.* **108**, 252.
Bose, S. K. 1980, *An Introduction to General Relativity*. Wiley Eastern, New Delhi, India.
Branch, D. 1998, *Ann. Rev. Astron. Astrophys.* **36**, 17.
Brandenberger, R. H. 1985, *Rev. Mod. Phys.* **57**, 1.
Brandenberger, R. H. 1987, *Int. J. Mod. Phys.* **A2**, 77.
Brecher, K. and Silk, J. 1969, *Ap. J.* **158**, 91.
Broadhurst, T. J., Ellis, R. S., Koo, D. C. and Szalay, A. S. 1990, *Nature* **343**, 726.
Burbidge, G. R. 1981, in *Tenth Texas Symposium on Relativistic Astrophysics*, eds. R. Ramaty and F. C. Jones, New York Academy of Sciences, p. 123.
Burbidge, G. R. and Burbidge, M. 1967, *Quasi-Stellar Objects*, W. H. Freeman & Co., San Francisco.
Burbidge, G. R., Burbidge, M., Fowler, W. A. and Hoyle, F. 1957, *Rev. Mod. Phys.* **29**, 547.
Calzetta, E. and Mazzitelli, F. D. 1990, Preprint GTCRG-90-3, Grupo de Teorias Cuanti Relativistas y Gravitacion, University of Buenos Aires.
Carr, B. 1988, *Nature* **334**, 650.
Carroll, S. M., Press, W. H. and Turner, E. L. 1992, *Ann. Rev. Astron. Astrophys*. **30**, 499.
Cesarsky, D. A., Moffet, A. T. and Pasachoff, J. M. 1973, *Ap. J. Lett.* **180**, L1.
Chandrasekhar, S. 1935, *Mon. Not. Roy. Astr. Soc.* **95**, 207.
Chandrasekhar, S. 1939, *An Introduction to the Study of Stellar Structure*, Dover Publications, New York.
Chandrasekhar, S. 1960, *Principles of Stellar Dynamics*, Dover Publications, New York.
Chernin, A. D. 1965, *Astr. Zh.* **42**, 1124.
Chernin, A. D. 1968, *Nature*, *Lond*. **220**, 250.
Chew, G. F. 1962, *S-Matrix Theory of Strong Interactions*, Benjamin, New York.
Chincarini, G. L., Giovanelli, R. and Haynes, M. P. 1983, *Ap. J.* **269**, 13.
Chincarini, G. and Rood, H. J. 1976, *Ap. J.* **206**, 30.
Clarke, C. J. S. 1993, *Analysis of Space-Time Singularities*, Cambridge University Press.
Coleman, S. 1988, *Nucl. Phys.* **B310**, 643.
Coleman, S. and Weinberg, E. 1973, *Phys. Rev*. **D7**, 788.
Conrath, B., Gautier, D. and Hornstein, J. 1982, *Saturn Meeting*, Tucson.
Crane, P. and Saslaw, W. C. 1986, *Ap. J.* **301**, 1.
Davies, P. C. W. 1973, *Mon. Not. Roy. Astr. Soc.* **161**, 1.
Davies, R. L. *et al*. 1987, *Ap. J.* **313**, 42, 15-D7; L-37, 27–49; **318**, 944, 91-C9.
Davis, M. 1990, *Nature*, *Lond*. **343**, 699.
D'Eath, P.D. 1996, *Supersymmetric Quantum Cosmology*, Cambridge University Press.
de Vaucouleurs, G. 1977, *Nature*, *Lond*. **266**, 125.

Dirac, P. A. M. 1975, *General Theory of Relativity*, John Wiley & Sons, reprinted 1996, Princeton University Press.
Djorgovski, S. and Spinrad, H. 1981, *Ap. J.* **251**, 417.
Dyson, F. J. 1979, *Rev. Mod. Phys*. **51**, 447.
Eardley, D., Liang, E. and Sachs, R. 1972, *J. Math. Phys.* **13**, 99.
Eddington, P. 1930, *Mon. Not. Roy. Astr. Soc.* **90**, 668.
Einstein, A. 1950, *The Meaning of Relativity*, 3rd edn., Princeton University Press, Princeton, N.J.
Eisenhart, L. P. 1926, *Riemannian Geometry*, Princeton University Press, Princeton, N.J.
Ellis, G. F. R. and Madsen, M. S. 1991, *Class. Quant. Grav.* **8**, 667.
Faddeev, L. D. and Popov, V. N. 1967, *Phys. Lett.* **25B**, 29.
Feynman, R. P. 1948, *Rev. Mod. Phys.* **20**, 367.
Feynman, R. P. and Hibbs, A. R. 1965, *Quantum Mechanics and Path Integrals*, McGraw-Hill Book Company, New York.
Freedman, W. L. *et al.* 1994, *Nature* **371**, 757.
Frenk, C. S., White, S. D. M., Davis, M. and Efstathiou, G. 1988, *Nature* **327**, 507.
Frenk, C. S., White, S. D. M., Efstathiou, G. and Davis, M. 1990, *Nature* **351**, 10.
Freund, P. G. O. 1986, *Introduction to Supersymmetry*, Cambridge University Press, Cambridge, England.
Friedman, I., Redfield, A. C., Schoen, B. and Harris, J. 1964, *Rev. Geophys*. **2**, 177.
Friedman, J. L. and Morris, M. S. 1990, *Nature* **343**, 409.
Friedmann, A. 1922, *Z. Phys.* **10**, 377.
Fukugita, M. and Hogan, C. J. 1990, *Nature* **347**, 120.
Futamase, T. and Maeda, K. 1989, *Phys. Rev.* **D39**, 399.
Futamase, T., Rothman, T. and Matzner, R. 1989, *Phys. Rev.* **D39**, 405.
Gamow, G. 1948, *Nature* **162**, 680.
Gautier, D. *et al*. 1981, *J. Geophys. Res.* **86**, 8713.
Gautier, D. and Owen, T. 1983, *Nature*, *Lond.* **302**, 215.
Geiss, J. and Reeves, H. 1972, *Astron. Astrophys.* **18**, 126.
Gell-Mann, M. and Hartle, J. B. 1990, in *Complexity, Entropy and the Physics of Information*, Santa Fe Institute for Studies in the Science of Complexity, Vol. IX, edited by W. H. Zurek, Addison Wesley.
Geroch, R. P. 1967, *Singularities in the Spacetime of General Relativity: Their Definition, Existence and Local Characterization*, Ph.D. thesis, Princeton University.
Gödel, K. 1949, *Rev. Mod. Phys.* **21**, 447.
Goode, S. W. and Wainwright, J. 1982, *Mon. Not. Roy. Astr. Soc.* **198**, 83.
Gregory, S. A. and Thompson, L. A. 1982, *Ap. J.* **286**, 422.
Gregory, S. A., Thompson, L. A. and Tifft, W. G. 1981, *Ap. J.* **243**, 411.
Grevesse, N. 1970, *Colloque de Liège*, **19**, 251.
Griest, K. and Silk, J. 1990, *Nature*, *Lond*. **343**, 222.
Gunn, J. E. 1978, in *Observational Cosmology*, eds. A. Maeder, L. Martinet and G. Tammann: Geneva Observatory, Geneva, Switzerland.
Gunn, J. E. and Oke, J. B. 1975, *Ap. J.* **195**, 255.
Gunn, J. E. and Tinsley, B. M. 1975, *Nature*, *Lond.* **257**, 454.
Guth, A. 1981, *Phys. Rev.* **D23**, 347.
Halliwell, J. J. and Hartle, J. B. 1990, *Phys. Rev.* **D41**, 1815.
Halliwell, J. J. and Louko, J. 1989a, *Phys. Rev.* **D39**, 2206.
Halliwell, J. J. and Louko, J. 1989b, *Phys. Rev.* **D40**, 1868.
Harrison, B. K., Thorne, K. S., Wakano, M. and Wheeler, J. A. 1965, *Gravitation*

Theory and Gravitational Collapse, University of Chicago Press, Chicago, USA.
Hartle, J. B. 1984, *Phys. Rev.* **D29**, 2730.
Hartle, J. B. 1986, *Prediction in Quantum Cosmology*, Lectures delivered at the 1986 Cargèse NATO Advanced Summer Institute, 'Gravitation and Astrophysics'.
Hartle, J. B. and Hawking, S. W. 1983, *Phys. Rev.* **D28**, 2960.
Hawking, S. W. 1975, *Comm. Math. Phys.* **43**, 199.
Hawking, S. W. and Ellis, G. F. R. 1973, *The Large-Scale Structure of Space-Time*, Cambridge University Press, Cambridge, England.
Hawking, S. W. and Penrose, R. 1970, *Proc. Roy. Soc.* **A314**, 529.
Heasley, J. and Milkey, R. 1978, *Ap. J.* **221**, 677.
Henry, J. P., Briel, U. G. and Böhringer, H. 1998, *Scientific American*, December.
Hewish, A., Bell, S. J., Pilkington, J. D. H., Scott, P. F. and Collins, R. A. 1968, *Nature*, *Lond*. **217**, 709.
Hirata, K. *et al*. 1987, *Phys. Rev. Lett.* **58**, 1490, 1798 (E).
Hodges, H. M. 1989, *Phys. Rev.* **D40**, 1798.
Hogan, C. J. 1990, *Nature*, *Lond.* **344**, 107.
Hogan, C. J., Kirshner, R. B. and Suntzeff, N. B. 1999, *Scientific American*, January.
Hoyle, F. 1948, *Mon. Not. Roy. Astr. Soc.* **108**, 72.
Hoyle, F. 1959, in *Paris Symposium on Radio Astronomy*, ed. R. N. Bracewell, Stanford University Press, Stanford, USA, p. 529.
Hoyle, F., Burbidge, G. and Narlikar, J. V. 2000, *A Different Approach to Cosmology*, Cambridge University Press.
Hoyle, F. and Narlikar, J. V. 1974, *Action at a Distance in Physics and Cosmology*, Freeman, San Francisco.
Huang, K. and Weinberg, S. 1970, *Phys. Rev. Lett.* **25**, 895.
Hubble, E. P. 1929, *Proc. Nat. Acad. Sci. USA* **15**, 169.
Hubble, E. P. 1934, *Ap. J.* **79**, 8.
Hubble, E. P. 1936, *Realm of the Nebulae*, Yale University Press, New Haven, Conn., USA.
Huchra, J., Davis, M., Latham, D. and Tonry, J. 1983, *Ap. J. Suppl*. **52**, 89.
Iben, I. Jr. 1969, *Ann. Phys.* **54**, 164.
Islam, J. N. 1977, *Quart. J. Roy. Astr. Soc.* **18**, 3.
Islam, J. N. 1979a, *Sky and Telescope* **57**, 13.
Islam, J. N. 1979b, *Vistas in Astronomy* **23**, 265.
Islam, J. N. 1983a, *The Ultimate Fate of the Universe*, Cambridge University Press, Cambridge, England.
Islam, J. N. 1983b, *Phys. Lett.* **97A**, 239.
Islam, J. N. 1984, *Endeavour*, *New Series* **8**, 32.
Islam, J. N. 1985, *Rotating Fields in General Relativity*, Cambridge University Press, Cambridge, England.
Islam, J. N. 1989, *Proc. Roy. Soc.* **A421**, 279.
Islam, J. N. 1993, *Progr. Theor. Phys.* **89**, 161.
Islam, J. N. 1994, *Found. Phys.* **24**, 593.
Islam, J. N. 2001a, 'An exact inflationary solution for a sixth degree potential', preprint.
Islam, J. N. 2001b, 'A brief overview of some aspects of astrophysics and cosmology', preprint, to appear in *Indian Journal of Physics*.
Islam, J. N. 2001c, 'Aspects of particle physics, gravitation and cosmology',

preprint, to appear in the *Proceedings of the Third World Academy of Sciences*.
Islam, J. N. and Munshi, N. I. 1990, 'On a Limit to the Cosmological Constant', unpublished.
Itzykson, C. and Zuber, J.-B. 1980, *Quantum and Field Theory*, McGraw-Hill.
Jacoby, G. H., Ciardullo, R. and Ford, H. C. 1990, *Ap. J.* **356**, 332.
Jefferts, K. B., Penzias, A. A. and Wilson, R. W. 1973, *Ap. J.* **179**, L57.
Joshi, P. S. and Narlikar, J. V. 1986, *Int. J. Mod. Phys.* **1**, 243.
Karachentsev, J. D. 1966, *Astrofizica* **2**, 81.
Kardashev, N. 1967, *Ap. J.* **150**, L135.
Kasner, E. 1921, *Am. J. Math.* **43**, 126, 130, 217.
Kearns, E., Kajita, T. and Totsuka, Y. 1999, *Scientific American*, August.
Khalatnikov, I. M. and Lifshitz, E. M. 1970, *Phys. Rev. Lett.* **24**, 76.
Kirshner, R. P., Oemler, A., Schechter, P. L. and Schectman, S. A. 1981, *Ap. J. Lett.* **248**, L57.
Kirzhnits, D. A. and Linde, A. D. 1972, *Phys. Lett.* **42B**, 471.
Kraan-Korteweg, R. C., Sandage, A. and Tammann, G. A. 1984, *Ap. J.* **283**, 24.
Kraft, R. 1961, *Ap. J.* **134**, 616.
Krauss, L. M. 1998, *Ap. J.* **501**, 461.
Krauss, L. M. 1999, *Scientific American*, January.
Krauss, L. M. and Starkman, G. D. 1999, *Scientific American*, November.
Kristian, J., Sandage, A. and Westphal, J. A. 1978, *Ap. J.* **221**, 383.
Kruskal, M. D. 1960, *Phys. Rev.* **119**, 1743.
Kunde, V. *et al.* 1982, *Ap. J.* **263**, 443.
Kung, J. H. and Brandenberger, R. 1989, *Phys. Rev.* **D40**, 2532.
La, D. and Steinhardt, P. J. 1989, *Phys. Rev. Lett.* **62**, 376.
La, D., Steinhardt, P. J. and Bertschinger, E. W. 1989, *Phys. Lett.* **B231**, 231.
Lambert, D. 1967, *Observatory* **960**, 199.
Landau, L. D. and Lifshitz, E. M. 1975, *The Classical Theory of Fields*, 4th English edn., Pergamon Press, Oxford.
Landsberg, P. T. and Park, D. 1975, *Proc. Roy. Soc. Lond.* **A346**, 485.
Lanzetta, K. M., Wolfe, A. M. and Turnshek, D. 1995, *Ap. J.* **440**, 435.
Lazarides, G. 1997, *Inflation*, Lectures given at the BCSPIN Summer School, Kathmandu, Nepal.
Lemaître, G. 1927, *Ann. Soc. Sci. Brux.* **A47**, 49.
Lemaître, G. 1931, *Mon. Not. Roy. Astron. Soc.* **91**, 483.
Liddle, A. R. *et al.* 1996, *Mon. Not. Roy. Astron. Soc.* **278**, 644; **281**, 531; **282**, 281.
Liddle, A. R. and Lyth, D. H. 2000, *Cosmological Inflation and Large-Scale Structure*, Cambridge University Press.
Lifshitz, E. M. and Khalatnikov, I. M. 1963, *Adv. Phys.* **12**, 185.
Lifshitz, E. M. and Khalatnikov, I. M. 1971, *Sov. Phys. JETP Lett.* **11**, 123.
Lilley, S. J. and Longair, M. S. 1984, *Mon. Not. Roy. Astr. Soc.* **211**, 833.
Lilley, S. J., Longair, M. S. and Allington-Smith, J. R. 1985, *Mon. Not. Roy. Astr. Soc.* **215**, 37.
Linde, A. D. 1982, *Phys. Lett*. **108B**, 389.
Linde, A. D. 1994, *Scientific American*, November.
Lindley, D. 1990a, *Nature, Lond.* **343**, 207.
Lindley, D. 1990b, *Nature, Lond.* **345**, 23.
Longair, M. S. 1978, in *Observational Cosmology*, eds. A. Maeder, L. Martinet and G. Tammann, Geneva Observatory, Geneva, Switzerland.
Longair, M. S. 1983, *Phys. Bull.* **34**, 106.
Longair, M. S. 1998, *Galaxy Formation*, Springer.

MacCallum, M. A. H. 1973, in *Cargese Lectures in Physics*, Vol. 6, ed. E. Schatzmann, Gordon and Breach, New York.
Marshak, R. E., Riazuddin and Ryan, C. P. 1969, *Theory of Weak Interactions in Particle Physics*, Wiley-Interscience, New York.
Mazenko, G., Unruh, W. and Wald, R. 1985, *Phys. Rev.* **D31**, 273.
Mazzitelli, I. 1979, *Astr. Astrophys.* **80**, 155.
McCrea, W. H. and Milne, E. A. 1934, *Q. J. Math.* **5**, 73, 76.
McIntosh, J. 1968, *Mon. Not. Roy. Astr. Soc.* **140**, 461.
Miller, J. C. and Pantano, O. 1989, *Phys. Rev.* **D40**, 1789.
Misner, C. W. 1969, *Phys. Rev. Lett.* **22**, 1071.
Misner, C. W., Thorne, K. S. and Wheeler, J. A. 1973, *Gravitation*. W. H. Freeman and Company, San Francisco.
Morris, M., Thorne, K. and Yurtsever, U. 1988, *Phys. Rev. Lett.* **61**, 1446.
Munshi, N. I. 1999, Ph.D. thesis, University of Chittagong.
Nakamura, T. T. and Suto, Y. 1995, *Ap. J.* **447**, L65.
Narlikar, J. V. 1979, *Mon. Not. Roy. Astr. Soc.* **183**, 159.
Narlikar, J. V. and Padmanabhan, T. 1983, *Phys. Rep*. **100**, 152.
Oke, J. B. and Sandage, A. 1968, *Ap. J.* **154**, 21.
Oldershaw, R. L. 1990, *Nature, Lond*. **346**, 800.
Oort, J. H. 1983, *Ann. Rev. Astron. Astrophys.* **21**, 373.
Orton, G. and Ingersoll, A. 1980, *J. Geophys. Res.* **85**, 5871.
Ostriker, J. and Tremaine, S. 1975, *Ap. J. Lett.* **202**, L113.
Pacher, T., Stein-Schabes, J. A. and Turner, M. S. 1987, *Phys. Rev.* **D36**, 1603.
Page, D. N. 1987, *Phys. Rev.* **D36**, 1607.
Page, D. N. and McKee, M. R. 1981, *Nature, Lond.* **291**, 44.
Pagel, B. 1984, *Philos. Trans. Roy. Soc. Lond.* **310A**, 245.
Pasachof, J. M. and Cesarsky, D. A. 1974, *Ap. J.* **193**, 65.
Peebles, P. J. E. 1966, *Phys. Rev. Lett.* **16**, 410.
Peebles, P. J. E. 1971, *Physical Cosmology*, Princeton University Press, Princeton, N.J.
Peebles, P. J. E. and Silk, J. 1988, *Nature, Lond.* **335**, 601.
Peebles, P. J. E. and Silk, J. 1990, *Nature, Lond.* **346**, 233.
Penzias, A. A. and Wilson, R. W. 1965, *Ap. J.* **142**, 419.
Perlmutter, S. *et al.* 1998, *Nature*, **391**, 51.
Petrosian, V. and Salpeter, E. E. 1968, *Ap. J.* **151**, 411.
Petrosian, V., Salpeter, E. E. and Szekeres, P. 1967, *Ap. J.* **147**, 1222.
Press, W. H. and Schechter, P. 1974, *Ap. J.* **187**, 452.
Press, W. H. and Spergel, D. N. 1989, *Physics Today*, March.
Rahaman, N. 1996, M.Sc. thesis, University of Dhaka.
Rahaman, N. and Rashid, A. M. H. 1996, unpublished.
Raychaudhuri, A. K. 1955, *Phys. Rev.* **98**, 1123.
Raychaudhuri, A. K. 1958, *Proc. Phys. Soc. Lond.* **72**, 263.
Raychaudhuri, A. K. 1979, *Theoretical Cosmology*, Oxford University Press. Oxford, England.
Raychaudhuri, A. K. 1998, *Phys. Rev. Lett.* **80**, 654; **81**, 5033.
Raychaudhuri, A. K. 1999, *Singularity Theorems Re-visited*, Lecture given at the tenth anniversary of the Inter-University Centre for Astronomy and Astrophysics, Pune, India.
Rees, M. J. 1969, *Observatory* **89**, 193.
Rees, M. J. 1987, in *Proceedings of the International Conference on Mathematical Physics*, ed. J. N. Islam, University of Chittagong, Bangladesh.
Rees, M. J. 1999, *Scientific American*, December.

Rindler, W. 1956, *Mon. Not. Roy. Astr. Soc.* **116**, 6.
Robertson, H. P. 1935, *Ap. J.* **82**, 248.
Robertson, H. P. 1936, *Ap. J.* **83**, 187.
Rogerson, J. B. Jr. and York, D. G. 1973, *Ap. J. Lett.* **186**, L95.
Rossi, G. C. and Testa, M. 1984, *Phys. Rev.* **D29**, 2997.
Ruiz, E. and Senovilla, J. M. M. 1992, *Phys. Rev.* **D45**, 1995.
Ryan, M. P. Jr. and Shepley, L. C. 1975, *Homogeneous Relativistic Cosmologies*, Princeton University Press, Princeton, N.J.
Ryder, L. 1996, *Quantum Field Theory*, Second edn, Cambridge University Press.
Saha, A. *et al.* 1995, *Ap. J.* **438**, 8.
Salam, A. 1987, in *Proceedings of the International Conference on Mathematical Physics*, ed. J. N. Islam, University of Chittagong, Bangladesh.
Salam, A. 1989, in *The New Physics*, ed. P. Davies, Cambridge University Press.
Salati, P. 1990, *Nature, Lond.* **346**, 221.
Sandage, A. 1968, *Observatory* **88**, 91.
Sandage, A. 1970, *Physics Today*, February.
Sandage, A. 1972a, *Ap. J.* **173**, 485.
Sandage, A. 1972b, *Ap. J.* **178**, 1.
Sandage, A. 1972c, *Quart. J. R. Astr. Soc.* **13**, 282.
Sandage, A. 1975a, in *Galaxies and the Universe*: eds. A. Sandage, M. Sandage and J. Kristian, University of Chicago Press, Chicago, p. 761.
Sandage, A. 1975b, *Ap. J.* **202**, 563.
Sandage, A. 1982, *Ap. J.* **252**, 553.
Sandage, A. 1987, *Proceedings of the IAU Symposium, Beijing, China*.
Sandage, A. and Hardy, E. 1973, *Ap. J.* **183**, 743.
Sandage, A., Katem, B. and Sandage, M. 1981, *Ap. J. Suppl.* **46**, 41.
Sandage, A. and Tammann, G. A. 1968, *Ap. J.* **151**, 531.
Sandage, A. and Tammann, G. A. 1969, *Ap. J.* **157**, 683.
Sandage, A. and Tammann, G. A. 1975, *Ap. J.* **197**, 265.
Sandage, A. and Tammann, G. A. 1983, in *Large Scale Structure of the Universe, Cosmology and Fundamental Physics*, First ESO-CERN Conference, eds. G. Setti and L. van Hove, Garching, Geneva, p. 127.
Sandage, A. and Tammann, G. A. 1986, in *Inner Space Outer Space*, eds. E. W. Kolb, M. S. Turner, D. Lindley, K. Olive and D. Steckel: University of Chicago Press, Chicago, p. 41.
Sandage, A. and Tammann, G. A. and Yahil, A. 1979, *Ap. J.* **232**, 352.
Saslaw, W. C. 1973, *Publ. Astr. Soc. Pacific* **85**, 5.
Saslaw, W. C. 2000, *The Distribution of the Galaxies: Gravitational Clustering in Cosmology*, Cambridge University Press.
Saslaw, W. C., Valtonen, M. J. and Aarseth, S. J. 1974, *Ap. J.* **190**, 253.
Schmidt, B. G. 1973, *Commun. Math. Phys.* **29**, 49.
Schneider, D. P., Schmidt, M. and Gunn, J. E. 1989, *Ap. J.* **98**, 1507, 1951.
Schramm, D. N. 1982 paper at Royal Society (London) meeting 11–12 March.
Schramm, D. N. and Wagoner, 1974, *Physics Today*, December.
Schücking, E. and Heckmann, J. 1958, 'World models', in *Onzieme Conseil de Physique Solvay*, Editions Stoops, Brussels, pp. 149–58.
Schwarzschild, B. 1989, *Physics Today*, March.
Schweber, S. S. 1961, *An Introduction to Relativistic Quantum Field Theory*, Row, Peterson and Co.
Seldner, M., Siebers, B., Groth, E. J. and Peebles, P. J. E. 1977, *Ap. J.* **82**, 249.
Senovilla, J. M. M. 1990, *Phys. Rev. Lett.* **64**, 2219.

Shellard, E. P. S. and Brandenberger, R. H. 1988, *Phys. Rev.* **D38**, 3610.
Smoluchowski, R. 1967, *Nature, Lond.* **215**, 691.
Spaenhauer, A. M. 1978, *Astron. Astrophys.* **65**, 313.
Spinrad, H. 1986, *Pub. A. S. P.* **98**, 269.
Steinhardt, P. J. 1990, *Nature, Lond.* **345**, 47.
Subramanian, K. and Padmanabhan, T. 1993, *Ap. J.* **410**, 482; **417**, 3.
Sunyaev, R. and Zel'dovich, Ya. B. 1980, *Ann. Rev. Astron. Astrophys.* **18**, 537.
Synge, J. L. 1937, *Proc. Lond. Math. Soc.* **43**, 376.
Synge, J. L. 1955, *Relativity: the General Theory*, North-Holland: Amsterdam.
Szafron, D. A. 1977, *J. Math. Phys.* **18**, 1673.
Szafron, D. A. and Wainwright, J. 1977, *J. Math. Phys.* **18**, 1668.
Szekeres, P. 1975, *Commun. Math. Phys.* **41**, 55.
Tammann, G. A., Yahil, A. and Sandage, A. 1979, *Ap. J.* **234**, 775.
Tarenghi, M., Tifft, W. G., Chincarini, G., Rood, H. J. and Thompson, L. A. 1979, *Ap. J.* **234**, 793.
Tayler, R. J. 1983, *Europhysics News* **14**, 1.
Taylor, J. C. 1976, *Gauge Theories of Weak Interactions*, Cambridge University Press.
Taylor, J. C. 1989, in *The New Physics*, ed. P. Davies, Cambridge University Press.
Thorne, K. S. 1967, *Ap. J.* **148**, 51.
Thorne, K. S. 1974, *Scientific American*, December.
Tinsley, B. M. 1977, *Physics Today*, June.
Tinsley, B. M. 1978, *Nature, Lond.* **273**, 208.
Trauger, J. T., Roesler, F. L., Carleton, N. P. and Traub, W. A. 1973, *Ap. J. Lett.* **184**, L137.
Turner, M. S. 1985, in *Proceedings of the Cargese School on Fundamental Physics and Cosmology*, eds. J. Audouze and J. Tran Thanh Van: Editions Frontièrs, Gif-Sur-Yvette.
Turner, M. S. *et al.* 1991, *Astron. J.* **103**, 1427.
Turok, N. 1989, *Phys. Rev. Lett.* **63**, 2625.
Ulrich, R. and Rood, R. 1973, *Nature Phys. Sci.* **241**, 111.
Van den Bergh, S. 1975, in *Galaxies and the Universe*, University of Chicago Press, Chicago.
Viana, P. T. P. and Liddle, A. R. 1996, *Mon. Not. Roy. Astron. Soc.* **278**, 644.
Vilenkin, A. 1990, *Nature, Lond.* **343**, 591.
Wagoner, R. V. 1973, *Ap. J.* **179**, 343.
Wagoner, R. V., Fowler, W. A. and Hoyle, F. 1967, *Ap. J.* **148**, 3.
Wainwright, J. 1979, *J. Phys. A* **12**, 2015.
Wainwright, J. 1981, *J. Phys. A* **14**, 1131.
Wainwright, J. and Goode, S. W. 1980, *Phys. Rev.* **D22**, 1906.
Wainwright, J., Ince, W. C. W. and Marshman, B. H. 1979, *Gen. Rel. Grav.* **10**, 259.
Wainwright, J. and Marshman, B. J. 1979, *Phys. Lett.* **72A**, 275.
Walker, A. G. 1936, *Proc. Lond. Math. Soc.* **42**, 90.
Weinberg, D. H., Ostriker, J. P. and Dekel, A. 1989, *Ap. J.* **336**, 9.
Weinberg, S. 1972, *Gravitation and Cosmology*, John Wiley and Sons, New York.
Weinberg, S. 1977, *The First Three Minutes*, André Deutsch, London; re-published in 1983 with an Afterword by Fontana Paperbacks.
Weinberg, S. 1989, *Rev. Mod. Phys.* **66**, 1.
Weinberg, S. 1999, *Scientific American*, December.
Weinreb, S. 1962, *Nature, Lond.* **195**, 367.
Weiss, N. 1989, *Phys. Rev.* **D39**, 1517.

Weyl, H. 1923, *Phys. Z.* **24**, 230.
White, S. D. M., Efstathiou, G. and Frenk, C. S. 1993, *Mon. Not. Roy. Astron. Soc.* **262**, 1023.
Wilson, R. W., Penzias, A. A., Jefferts, K. B. and Solomon, P. R. 1973, *Ap. J. Lett.* **179**, L107.
Zel'dovich, Ya. B. 1968, *Uspekhi* **95**, 209.
Zel'dovich, Ya. B., Einasto, J. and Shandarin, S. F. 1982, *Nature, Lond.* **300**, 407.

Index

absolute bolometric magnitude 91
absolute luminosity 72
age of the universe 5, 8, 61, 63, 65
angular diameter 88
anisotropic model 121
anti-de Sitter space 99
apparent bolometric magnitude 91
apparent luminosity 72

background neutrinos 87
background photons 87
Bianchi identity 14
black-hole evaporation 215
bolometric magnitude 80
broken symmetry 141

Cepheid variables 78
closed universe 5
clusters of galaxies 1
colour index 79
comoving coordinates 45
conformal fluctuations 206
cosmic background radiation 6, 9, 129, 218
cosmological constant 94–98, 102, 209
  limits to 100
Cosmological Principle 3, 37
covariant differentiation 13
critical density 7, 63, 64

deceleration parameter xi, 8, 50, 76, 80, 87, 92
density parameter 8, 64, 166
de Sitter group 96
de Sitter model 96, 97, 100
deuterium 134, 158
distance modulus 92
dynamical friction 85

early universe 128–133
Eddington–Lemaitre model 97
effective potential 181
Einstein equations 15
Einstein–de Sitter model 66
Einstein tensor 15
elementary particles 136
energy–momentum tensor 15
event horizon 4, 75

flatness problem 10, 166
foam structure of space-time 209
Friedmann models 5, 60–75
future of the universe 211–219

galactic black holes 213
galaxies, cluster of 1
galaxies, recession of 2
Gamow formula 216
general relativity, summary of 12
geodesics 16, 31, 39
globular clusters 89
Grand Unified Theories 140, 168, 181

Hamiltonian formalism 191
helium 131–133, 153, 158, 159
Higgs fields 168, 169
homogeneous cosmologies 113
homogeneous universe 3, 37, 41
horizon problem 10, 167
hot universe 147
Hubble flow 88
Hubble's constant 8, 76–79, 89, 108, 229
Hubble's law 2, 4, 6, 49
Hubble time 9, 61

inflationary models 166–178
inhomogeneous cosmologies 126
isometry 19
isotropic universe 3, 37

Jupiter, abundance of elements in 161–163

Killing vectors 18–21, 41, 53–59

Lemaitre models 96
luminosity distance 72

Malmquist bias 89
mass-energy conservation 16, 52
mixmaster singularity 125
monopoles 168, 181, 182
$m(z)$ test 88

neutrinos 87, 145, 146
neutrino temperature 108, 146
neutron abundance 133
nucleosynthesis 132, 153

observational cosmology 76–93
open universe 5
oscillatory approach to singularities 122

particle horizon 73, 74
path integrals 202
Planck length 206

quantum cosmology 11, 189, 234–237

recombination 134
red-shift 2, 48
relativistic hydrodynamics 115
Ricci tensor 14
Riemann tensor 14
Robertson–Walker metric 37–40, 42–52
  rigorous derivation of 53–59

scale factor of the universe 3, 7
Scott effect 83
selection effects 88
singularity theorem 120
smoothness problem 10, 166
space-time singularity 5, 112, 117, 118
spontaneous symmetry breaking 168
stability of matter 217
standard candles 80
standard model 9, 10
stars, death of 211
Steady State Theory 96
superclusters 1, 88
superspace 202

temperature of early universe 143

universe,
  definition of 2
  closed 5
  early 128–131
  open 5
  very early 135, 166–173

Weyl's postulate 38
Wheeler–de Witt equation 201, 234–237

Yang–Mills field 205