OXFORD MASTER SERIES IN PARTICLE PHYSICS,
ASTROPHYSICS, AND COSMOLOGY

Relativity, Gravitation and Cosmology

A Basic Introduction

Ta-Pei Cheng

OXFORD MASTER SERIES IN PARTICLE PHYSICS, ASTROPHYSICS, AND COSMOLOGY

OXFORD MASTER SERIES IN PHYSICS

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

CONDENSED MATTER PHYSICS

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

ATOMIC, OPTICAL, AND LASER PHYSICS

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

PARTICLE PHYSICS, ASTROPHYSICS, AND COSMOLOGY

10. D. H. Perkins: *Particle astrophysics*
11. T. P. Cheng: *Relativity, gravitation, and cosmology*

STATISTICAL, COMPUTATIONAL, AND THEORETICAL PHYSICS

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

Relativity, Gravitation, and Cosmology

A basic introduction

TA-PEI CHENG

University of Missouri—St. Louis

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford OX2 6DP

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

Oxford New York

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

With offices in

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

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

First published 2005

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

Library of Congress Cataloging-in-Publication Data

Cheng, Ta-Pei.
Relativity, gravitation, and cosmology: a basic introduction /
Ta-Pei Cheng.
p. cm.—(Oxford master series in physics; no. 11)
Includes bibliographical references and index.
ISBN 0-19-852956-2 (alk. paper)—ISBN 0-19-852957-0 (pbk.
: alk. paper)
1. General relativity (Physics)—Textbooks. 2. Space and time.
3. Gravity. 4. Cosmology. I. Title. II. Series: Oxford master
series in physics; 11.
QC173.6.C4724 2005
530.11—dc22
2004019733

Typeset by Newgen Imaging Systems (P) Ltd., Chennai, India

Printed in Great Britain
on acid-free paper by Antony Rowe, Chippenham

ISBN 0 19 852956 2 (Hbk)
ISBN 0 19 852957 0 (Pbk)

10 9 8 7 6 5 4 3 2 1

Preface

It seems a reasonable expectation that every student receiving a university degree in physics will have had a course in one of the most important developments in modern physics: Einstein's general theory of relativity. Also, given the exciting discoveries in astrophysics and cosmology of recent years, it is highly desirable to have an introductory course whereby such subjects can be presented in their proper framework. Again, this is general relativity (GR).

Nevertheless, a GR course has not been commonly available to undergraduates, or even for that matter, to graduate students who do not specialize in GR or field theory. One of the reasons, in my view, is the insufficient number of suitable textbooks that introduce the subject with an emphasis on physical examples and simple applications without the full tensor apparatus from the very beginning. There are many excellent graduate GR books; there are equally many excellent "popular" books that describe Einstein's theory of gravitation and cosmology at the qualitative level; and there are not enough books in between. I am hopeful that this book will be a useful addition at this intermediate level. The goal is to provide a textbook that even an instructor who is not a relativist can teach from. It is also intended that other experienced physics readers who have not had a chance to learn GR can use the book to study the subject on their own. As explained below, this book has features that will make such an independent study particularly feasible.

Students should have had the usual math preparation at the calculus level, plus some familiarity with matrices, and the physics preparation of courses on mechanics and on electromagnetism where differential equations of Maxwell's theory are presented. Some exposure to special relativity as part of an introductory modern physics course will also be helpful, even though no prior knowledge of special relativity will be assumed. Part I of this book concentrates on the metric description of spacetime: first, the flat geometry as in special relativity, and then curved ones for general relativity. Here I discuss the equation of motion in Einstein's theory, and many of its applications: the three classical tests, black holes, and gravitational lensing, etc. Part II contains three chapters on cosmology. Besides the basic equations describing a homogeneous and isotropic universe, I present a careful treatment of distance and time in an expanding universe with a space that may be curved. The final chapter on cosmology, Chapter 9 provides an elementary discussion of the inflationary model of the big bang, as well as the recent discovery that the expansion of our universe is accelerating, implying the existence of a "dark energy." The tensor formulation of relativity is introduced in Part III. After presenting special relativity in a manifestly covariant formalism, we discuss covariant differentiation, parallel transport, and curvature tensor for a curved space. Chapter 12 contains the full tensor formulation of GR, including the Einstein's field equation and its solutions for various simple situations. The subject of gravitational waves can be found in the concluding chapter.

The emphasis of the book is pedagogical. The necessary mathematics will be introduced gradually. Tensor calculus is relegated to the last part of the book. Discussion of curved surfaces, especially the familiar example of a spherical surface, precedes that of curved higher dimensional spaces. Parts I and II present the metric description of spacetime. Many applications (including cosmology) can already be discussed at this more accessible level; students can reach these interesting results without having to struggle through the full tensor formulation, which is presented in Part III of the book. A few other pedagogical devices are also deployed:

- a **bullet list** of topical headings at the beginning of each chapter serves as the "chapter abstracts," giving the reader a foretaste of upcoming material;
- matter in marked **boxes** are calculation details, peripheral topics, historical tit-bits that can be skipped over depending on the reader's interest;
- **Review questions** at the end of each chapter should help beginning students to formulate questions on the key elements of the chapter[1]; brief answers to these questions are provided at the back of the book;
- **Solutions to selected problems** at the end of the book also contains some extra material that can be studied with techniques already presented in the text.

Given this order of presentation, with the more interesting applications coming before the difficult mathematical formalism, it is hoped that the book can be rather versatile in terms of how it can be used. Here are some of the possibilities:

1. Parts I and II should be suitable for an undergraduate course. The tensor formulation in Part III can then be used as extracurricular material for instructors to refer to, and for interested students to explore on their own. Much of the intermediate steps being given and more difficult problems having their solutions provided, this section can, in principle, be used as self-study material by a particularly motivated undergraduate.
2. The whole book can be used for a senior-undergraduate/beginning-graduate course. To fit into a one-semester course, one may have to leave some applications and illustrative examples to students as self-study topics.
3. The book is also suitable as a supplemental text: for an astronomy undergraduate course on cosmology, to provide a more detailed discussion of GR; for a regular advanced GR and cosmology course, to ease the transition for those graduate students not having had a thorough preparation in the relevant area.
4. The book is written keeping in mind readers doing independent study of the subject. The mathematical accessibility, and the various "pedagogical devices" (chapter headings, review questions, and worked-out solutions, etc.) should make it practical for an interested reader to use the book to study GR and cosmology on his or her own.

An updated list of corrections to the book can be found at the website `http://www.umsl.edu/~tpcheng/grbook.html`

[1] We find that the practice of frequent quizzes based on these review questions are an effective means to make sure that each member is keeping up with the progress of the class.

Acknowledgments

This book is based on the lecture notes of a course I taught for several years at the University of Missouri—St. Louis. Critical reaction from the students has been very helpful. Daisuke Takeshita, and also Michael Cone, provided me with detailed comments. My colleague Ricardo Flores has been very generous in answering my questions—be they in cosmology or computer typesetting. The painstaking task of doing all the line-drawing figures was carried out by Cindy Bertram. My editor Sonke Adlung at OUP has given me much support and useful advice. He arranged to have the manuscript reviewed by scholars who provided many suggestions for improvements. To all of them I am much indebted. Finally, I am grateful to my wife Leslie for her patient understanding during the rather lengthy period that it took me to complete this project.

Additional acknowledgment: I would like to express my gratitude to Professor Eric Sheldon. He was kind enough to read over the entire book and made numerous suggestions for editorial improvements, which were adopted in the new printings of this book.

St. Louis *T.P.C.*

This book is dedicated to
Professor Ling-Fong Li of Carnegie Mellon University
for more than 30 years' friendship and enlightenment

Contents

Part II COSMOLOGY

Part III RELATIVITY Full Tensor Formulation

This page intentionally left blank

RELATIVITY
Metric Description of Spacetime

Part I

This page intentionally left blank

Introduction and overview

1

- Relativity means that physically it is impossible to detect absolute motion. This can be stated as a symmetry in physics: physics equations are unchanged under coordinate transformations.
- Special relativity (SR) is the symmetry with respect to coordinate transformations among inertial frames, general relativity (GR) among more general frames, including the accelerating coordinate systems.
- The equivalence between the physics due to acceleration and to gravity means that GR is also the relativistic theory of gravitation, and SR is valid only in the absence of gravity.
- Einstein's motivations to develop GR are reviewed, and his basic idea of curved spacetime as the gravitation field is outlined.
- Relativity represents a new understanding of space and time. In SR we first learn that time is also a frame-dependent coordinate; the arena for physical phenomena is the four dimensional spacetime. GR interprets gravity as the structure of this spacetime. Ultimately, according to Einstein, space and time have no independent existence: they express relation and causal structure of physics processes in the world.
- The proper framework for cosmology is GR. The solution of the GR field equation describes the whole universe because it describes the whole spacetime.
- The outline of our presentation: Part I concentrates on the description of spacetime by the metric function. From this we can discuss many GR applications, including the study of cosmology in Part II. Only in Part III do we introduce the full tensor formulation of the GR field equations and the ways to solve them.

Einstein's general theory of relativity is a classical field theory of gravitation. It encompasses, and goes beyond, Newton's theory, which is valid only for particles moving with slow velocity (compared to the speed of light) in a weak and static gravitational field. Although the effects of general relativity (GR) are often small in the terrestrial and solar domains, its predictions have been accurately verified whenever high precision observations can be performed. Notably we have the three classical tests of GR: the precession of a planet's perihelion, the bending of star light by the sun, and redshift of light's frequency in a gravitational field. When it comes to situations involving strong gravity, such as compact stellar objects and cosmology, the use of GR is indispensable. Einstein's theory predicted the existence of black holes, where the gravity is so strong that even light cannot escape from them. We must also use GR for situations involving time-dependent gravitational fields as in emission and propagation of gravitational waves. The existence of gravitational waves as predicted by GR has been verified by observing the rate of energy loss, due to the emission of gravitational radiation, in a relativistic binary pulsar system. GR can naturally accommodate the possibility of a constant "vacuum energy density" giving rise to a repulsive gravitational force. Such an agent is the key ingredient of modern cosmological theories of the big bang (the inflationary cosmology) and of the accelerating universe (having a dark energy).

Creating new theories for the phenomena that are not easily observed on earth poses great challenges. We cannot repeat the steps that led to the formulation of Maxwell's theory of electromagnetism, as there are not many experimental results one can use to deduce their theoretical content. What Einstein pioneered was the elegant approach of using physics symmetries as a guide to the new theories that would be relevant to the yet-to-be-explored realms. As we shall explain below, relativity is a coordinate symmetry. Symmetry imposes restriction on the equations of physics. The condition that the new theory should be reduced to known physics in the appropriate limit often narrows it further down to a very few possibilities. The symmetry Einstein used for this purpose is the coordinate symmetries of relativity, and the guiding principle in the formulation of GR is the "principle of general covariance." In Section 1.1 we shall explain the meaning of a symmetry in physics, as well as present a brief historical account of the formulation of relativity as a coordinate symmetry. In Section 1.2 we discuss the motivations that led Einstein to his geometric view of gravitation that was GR.

Besides being a theory of gravitation, GR, also provides us with a new understanding of space and time. Starting with special relativity (SR), we learnt that time is not absolute. Just like spatial coordinates, it depends on the reference frame as defined by an observer. This leads to the perspective of viewing physical events as taking place in a 4D continuum, called the spacetime. Einstein went further in GR by showing that the geometry of this spacetime was just the phenomenon of gravitation and was thus determined by the matter and energy distribution. Ultimately, this solidifies the idea that space and time do not have an independent existence; they are nothing but mirroring the relations among physical events taking place in the world.

General relativity is a classical theory because it does not take into account quantum effects. GR being a theory of space and time means that any viable theory of quantum gravity must also offer a quantum description of space and time. Although quantum gravity[1] is beyond the scope of this book, we should nevertheless mention that current research shows that such a quantum theory has rich enough structure as to be the unified theory of all matter and interactions (gravitation, strong and electroweak, etc.). Thus the quantum generalization of GR should be **the** fundamental theory in physics.

In this introductory chapter, we shall put forward several "big motifs" of relativity, without much detailed explanation. Our purpose is to provide the reader with an overview of the subject—a roadmap, so to speak. It is hoped that, proceeding along the subsequent chapters, the reader will have occasion to refer back to this introduction, to see how various themes are substantiated.

[1] Currently the most developed study of quantum gravity is the string theory. For a recent textbook exposition see (Zwiebach, 2004).

1.1 Relativity as a coordinate symmetry

We are all familiar with the experience of sitting in a train, and not able to "feel" the speed of the train when it is moving with a constant velocity, and, when observing a passing train on a nearby track, find it difficult to tell which train is actually in motion. This can be interpreted as saying that no physical measurement can detect the absolute motion of an inertial frame. Thus we have the basic concept of **relativity**, stating that only relative motion is measurable in physics.

In this example, the passenger is an observer who determines a set of coordinates (i.e. rulers and clocks). What this observer measures is the physics with respect to this coordinate frame. The expression "the physics with respect to different coordinate systems" just means "the physics as deduced by different observers." Physics should be independent of coordinates. Such a statement proclaims a **symmetry in physics**: Physics laws remain the same (i.e. physics equations keep the same form) under some **symmetry transformation**, which changes certain conditions, for example, the coordinates. The invariance of physics laws under coordinate transformation is called **symmetry of relativity**. This coordinate symmetry can equivalently be stated as the impossibility of any physical measurement to detect a coordinate change. Namely, if the physics remains the same in all coordinates, then no experiment can reveal which coordinate system one is in, just as the passenger cannot detect the train's constant-velocity motion.

Rotational symmetry is a familiar example of coordinate symmetry. Physics equations are unchanged when written in different coordinate systems that are related to each other by rotations. Rotational symmetry says that it does not matter whether we do an experiment facing north or facing southwest. After discounting any peculiar local conditions, we should discover the same physics laws in both directions. Equivalently, no internal physical measurement can detect the orientation of a laboratory. The orientation of a coordinate frame is not absolute.

1.1.1 From Newtonian relativity to aether

Inertial frames of reference are the coordinate systems in which, according to Newton's First Law, a particle will, if no external force acts on it, continue its state of motion with constant velocity (including the state of rest). Galileo and Newton taught us that the physics description would be the simplest when given in these coordinate systems. The First Law provides us the definition of an inertial system (also called Galilean frames of reference). Its implicit message that such coordinate systems exist is its physical content. Nevertheless, the First Law does not specify which are the inertial frames in the physical universe. It is an empirical fact[2] that these are the frames moving at constant velocities with respect to the fixed stars—distant galaxies, or, another type of distant matter, the cosmic microwave background (CMB) radiation (see Section 8.5). There are infinite sets of such frames: differing by their relative orientation, displacement, and relative motion with constant velocities. For simplicity we shall ignore the transformations of rotation and displacement of coordinate origin, and concentrate on the relation among the rectilinear moving coordinates—the **boost** transformation.

[2] That there should be a physical explanation why the distant matter defines the inertial frames was first emphasized by Bishop George Berkeley in the eighteenth century, and by Ernst Mach in the nineteenth. A brief discussion of Mach's principle can be found in Box 1.1.

Physics equations in classical mechanics are invariant under such boost transformations. Namely, no mechanical measurement can detect the moving spatial coordinates. The familiar example of not being able to feel the speed of a moving train cited at the beginning of this section is a simple illustration of this **principle of Newtonian relativity**: "physics laws (classical mechanics) are the same in all inertial frames of reference." In this sense, there is no absolute rest frame in Newtonian mechanics. The situation changed when electromagnetism was included. Maxwell showed a light speed being given by the static parameters of electromagnetism. Apparently there is only one speed of light c regardless of whether the observer is moving or not. Before Einstein, just about everyone took it to mean that the Maxwell's equations were valid only in the rest frame of the **aether**, the purported medium for electromagnetic wave propagation. In effect this reintroduced into physics the notion of absolute space (the aether frame).

Also, in Newtonian mechanics the notion of time is taken to be absolute, as the passage of time is perceived to be the same in all coordinates.

1.1.2 Einsteinian relativity

It is in this context that one must appreciate Einstein's revolutionary proposal: All motions are relative and there is no need for concepts such as absolute space. Maxwell's equations are valid in every inertial coordinate system.[3] There is no aether. Light has the peculiar property of propagating with the same speed c in all (moving) coordinate systems—as confirmed by the Michelson–Morley experiment.[4] Furthermore, the constancy of the light speed implies that, as Einstein would show, there is no absolute time.

Einstein generalized the Newtonian relativity in two stages:

1905 Covariance of physics laws under boost transformations were generalized from Newtonian mechanics to include electromagnetism. Namely, the laws of electricity and magnetism, as well as mechanics, are unchanged under the coordinate transformations that connect different inertial frames of reference. Einstein emphasized that this generalization implied a new kinematics: not only space but also time measurements are coordinate dependent. It is called the principle of **special relativity** because we are still restricted to the special class of coordinates: the inertial frames of reference.

1915 The generalization is carried out further; **General relativity** is the physics symmetry allowing for more general coordinates, including the accelerating frames as well. Based on the empirical observation that the effect of an accelerating frame and gravity is the same, GR is the field theory of gravitation; SR is special because it is valid only in the absence of gravity. GR describes gravity as the curved spacetime, which, in SR, is flat.

To recapitulate, relativity is a coordinate symmetry. It is the statement that physics laws are the same in different coordinate systems. Thus, physically it is impossible to detect absolute motion and orientation because physics laws are unchanged under coordinate transformations. For SR, these are the transformations among Galilean frames of reference (where gravity is absent); for GR, among more general frames, including the accelerating coordinate systems.

[3] While emphasizing Einstein's role, we must also point out the important contribution to SR by Henri Poincaré. In fact the full Lorentz transformation was originally written down by Poincaré (who named it in Lorentz's honor). Poincaré was the first one to emphasize the view of relativity as a physics symmetry. For an accessible account of Poincaré's contribution, see Logunov (2001).

[4] Michelson and Morley, using a Michelson interferometer, set out to measure a possible difference in light speeds along and transverse to the orbit motion of the earth around the sun. Their null result confirmed the notion that light speed was the same in different inertial frames.

1.1.3 Coordinate symmetry transformations

Relativity is the symmetry describing the covariance of the physics equation (i.e. invariance of the equation form) under coordinate transformations. We need to distinguish among several classes of transformations:

Galilean transformation. In classical (nonrelativistic) mechanics, inertial frames are related to each other by this transformation. Thus, by Newtonian relativity, we mean that laws of Newtonian mechanics are covariant under Galilean transformations. From the modern perspective, Galilean transformations such as $t' = t$ are valid only when the relative velocity is negligibly small compared to c.

Lorentz transformation. As revealed by SR, the transformation rule connecting all the inertial frames, valid for all relative speed $< c$, is the Lorentz transformation. Namely, Galilean is the low-speed approximation of Lorentz transformation. Maxwell's equations are first discovered to possess this symmetry—they are covariant under the Lorentz transformation. It then follows that Newtonian (nonrelativistic) mechanics must be modified so that the relativistic mechanics, valid for particles having arbitrary speed up to c, can also have this Lorentz symmetry.

General coordinate transformation. The principle that physics equations should be covariant under the general transformations that connect different coordinate frames, including accelerating frames, is GR. Such a general symmetry principle is called the **principle of general covariance**. This is the basic principle guiding the construction of the relativistic theory of gravitation.

Thus, in GR, all sorts of coordinates are allowed—there is a "democracy of coordinate systems." All sorts of coordinate transformations can be used. But the most fruitful way of viewing the transformations in GR is that they are local (i.e. an independent one at every space–time point) Lorentz transformations, which in the low-velocity limit are Galilean transformations.

1.1.4 New kinematics and dynamics

Einstein's formulation of the relativity principle involves a sweeping change of kinematics: not only space, but also the time measurements, may differ in different inertial frames. Space and time are on equal footing as coordinates of a reference system. We can represent space and time coordinates as the four components of a (spacetime) position vector x^μ ($\mu = 0, 1, 2, 3$), with x^0 being the time component, and the transformation for coordinate differentials is now represented by a 4×4 matrix $[\mathbf{A}]$,

$$dx^\mu \rightarrow dx'^\mu = \sum_\nu [\mathbf{A}]^\mu_\nu dx^\nu, \qquad (1.1)$$

just as rotational coordinate transformation is represented by a 3×3 matrix. The Galilean and Lorentz transformations are **linear** transformations; that is, the transformation matrix elements do not themselves depend on the coordinates $[\mathbf{A}] \neq [\mathbf{A}(x)]$. The transformation matrix being a constant with respect to the coordinates means that one makes the **same** transformation at **every** coordinate point. We call this a **global transformation**. By contrast, the general coordinate transformations are **nonlinear** transformations.

Recall, for example, the transformation to an accelerating frame, $x \to x' = x + vt + at^2/2$, is nonlinear in the time coordinate. Here the transformations are coordinate-dependent, $[\mathbf{A}] = [\mathbf{A}(x)]$—a different transformation for each coordinate space–time point. We call this a **local transformation**, or a **gauge transformation**. Global symmetry leads to kinematic restrictions, while local symmetry dictates dynamics as well. As we shall see, the general coordinate symmetry (GR) leads to a dynamical theory of gravitation.[5]

1.2 GR as a gravitational field theory

The problem of noninertial frames of reference is intimately tied to the physics of gravity. In fact, the inertial frames of reference should properly be defined as the reference frames having no gravity. GR, which includes the consideration of accelerating coordinate systems, represents a new theory of gravitation.

The development of this new theory is rather unique in the history of physics: it was not prompted by any obvious failure (crisis) of Newton's theory, but resulted from the theoretical research, "pure thought," of one person—Albert Einstein. Someone put it this way: "Einstein just stared at his own navel, and came up with general relativity."[6]

1.2.1 Einstein's motivations for the general theory

If not prompted by experimental crisis, what were Einstein's motivations in his search for this new theory? From his published papers,[7] one can infer several **interconnected** motivations (Uhlenbeck, 1968):

1. To have a relativistic theory of gravitation. The Newtonian theory of gravitation is not compatible with the principle of (special) relativity as it requires the concept of "action-at-a-distance" force, which implies instantaneous transmission of signals.
2. To have a deeper understanding of the empirically observed equality between inertial mass and gravitational mass.
3. "Space is not a thing." Einstein phrased his conviction that physics laws **should not** depend on reference frames, which express the relationship among physical processes in the world and do not have independent existence.

Comments on this list of motivations

1. The Newtonian theory is nonrelativistic. Recall that Newton's theory of gravitation resembles Coulomb's law of electrostatics. They are static field theories with no field propagation. Eventually, the electromagnetic theory is formulated as a dynamical field theory. The source acts on the test charge not through the instantaneous action-at-a-distance type of force, but instead by the creation of electromagnetic fields which propagate out with a finite speed, the speed of light c. Thus the problem is how to formulate a field theory of gravitation with physical influence propagating at finite speed. More broadly speaking, one would like to have a new theory of gravity in which space and time are treated on more equal footing.

[5] Following Einstein's seminal work, physicists learned to apply the local symmetry idea also to the internal charge–space coordinates. In this way, electromagnetism as well as other fundamental interactions among elementary particles (strong and weak interactions) can all be understood as manifestation of local gauge symmetries. For respective references of gauge theory in general and GR as a gauge theory in particular, see for example (Cheng and Li, 1988 and 2000).

[6] The reader of course should not take this description to imply that the discovery was in any sense straightforward and logically self-evident. In fact, it took Einstein close to 10 years of difficult research, with many false detours, to arrive at his final formulation. To find the right mathematics of Riemannian geometry, he was helped by his friend and collaborator Marcel Grossmann.

[7] Einstein's classical papers in English translation may be found in the collected work published by Princeton University Press (Einstein, 1989). A less complete, but more readily available, collection may be found in (Einstein *et al.*, 1952).

2. In the course of writing a review paper on relativity in 1907 Einstein recalled the fundamental experimental result (almost forgotten since Newton's days) that the **gravitational mass** and the **inertial mass** are equal

$$m_{\mathrm{G}} = m_{\mathrm{I}}. \tag{1.2}$$

This is the essence of Galileo's observation in the famous "Leaning Tower experiment": all objects fall with the **same** acceleration. Inserting the gravitational force $m_{\mathrm{G}}g$ (where g is the gravitational acceleration) into Newton's Second Law $F = m_{\mathrm{I}}a$,

$$m_{\mathrm{G}}g = m_{\mathrm{I}}a, \tag{1.3}$$

we see that the empirical result:

$$a = g \qquad \text{same for all objects} \tag{1.4}$$

leads to the conclusion in (1.2). This equality $m_{\mathrm{I}} = m_{\mathrm{G}}$ is rather remarkable. While inertial mass m_{I} is the response of an object to all forces as it appears in $F = m_{\mathrm{I}}a$, the gravitational mass m_{G} is the response to (as well as the source of) a specific force: gravity—we can think m_{G} as the "gravitational charge" of an object. Viewed this way, we see the unique nature of gravitational force. No other fundamental force has this property of its response, the acceleration as shown in (1.4), being independent from any attribute of the test particle. On the other hand, such a property reminds us of the "fictitious forces," for example, centrifugal and Coriolis forces, etc.; the presence of such forces are usually attributed to a "bad choice" of frames (i.e. accelerating frames of reference). To highlight the importance of this experimental fact, Einstein elevated this equality (1.2) into **the equivalence principle** (EP):

$$\begin{pmatrix} \text{an inertial frame with gravity "}\mathbf{g}\text{"} \\ \text{is equivalent to} \\ \text{an accelerated frame with an acceleration of "}-\mathbf{g}\text{"} \end{pmatrix}.$$

This means that gravity and accelerated motion are indistinguishable. Once gravity is included in this framework, all frames of reference, whether in constant or accelerated motion, are now on equal footing. All coordinate transformations can be taken into consideration at the same time. Furthermore, with the problem stated in this way, Einstein was able to generalize this equivalence beyond mechanics. By considering the various links between gravity and accelerated motion, Einstein came up with the idea that gravity can cause the fabric of space (and time) to warp. Namely, the shape of space responds to the matter in the environment.

3. Einstein was dissatisfied with the prevailing concept of space. SR confirms the validity of the principle of special relativity: physics is the same in every Galilean frame of reference. But as soon one attempts to describe physical phenomena from a reference frame in acceleration with respect to an inertial frame, the laws of physics change and become more complicated because of the presence of the fictitious inertial forces. This is particularly troublesome from the viewpoint of relative motion, since one could identify either frame as the accelerating frame. (The example known as Mach's paradox is discussed in Box 1.1.) The presence of the inertial force is associated with the choice of a noninertial coordinate system. Such coordinate-dependent phenomena can be thought as brought about by space itself. Namely, space behaves as if it is

the source of the inertial forces. Newton was compelled thus to postulate the existence of **absolute space**, as the origin of these coordinate-dependent forces. The unsatisfactory feature of such an explanation is that, while absolute space is supposed to have an independent existence, yet no object can act on this entity. Being strongly influenced by the teaching of Ernst Mach (Box 1.1), Einstein emphasized that reference frames were human construct and true physics laws should be independent of coordinate frames. Space and time should not be like a stage upon which physical events take place, and thus have an existence even in the absence of physical interactions. In Mach's and Einstein's view, space and time are nothing but expressing relationships among physical processes in the world—"space is not a thing." Such considerations, together with the idea of the principle of equivalence between gravitation and inertial forces, led Einstein to the belief that the laws of physics should have the same form in all reference forms, thus abolishing the concept of space as a thing. If one knows the laws of physics in an inertial frame of reference having a gravitational field, and carries out a transformation to a frame accelerating with respect to the first one, then the effect of acceleration must be the same as that due to gravity in the first. In Chapter 3, we shall provide several examples showing how to extract physical consequences from applications of this EP.

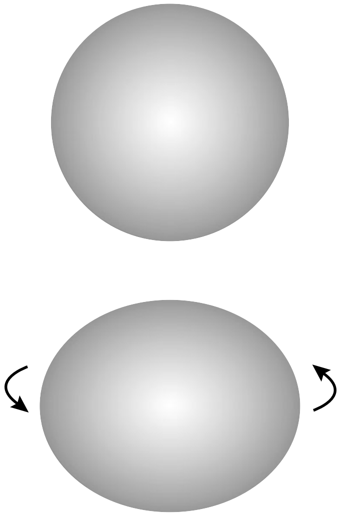

Fig. 1.1 **Mach's paradox**: Two identical elastic spheres, one at rest, and the other rotating, in an inertial frame of reference. The rotating sphere is observed to bulge out in the equatorial region, taking on an ellipsoidal shape. (For proper consideration, the two spheres should be separated by a distance much larger than their size.)

Box 1.1 Mach's principle

At the beginning of his 1916 paper on general relativity, Einstein discussed **Mach's paradox** (Fig. 1.1) to illustrate the unsatisfactory nature of Newton's conception of space as an active agent. Consider two identical elastic spheres separated by a distance much larger than their size. One is at rest, and the other rotating around the axis joining these two spheres in an inertial frame of reference. The rotating body takes on the shape of an ellipsoid. Yet if the spheres are alone in the world, each can be regarded as being in rotation with respect to the other. Thus there should be no reason for dissimilarity in shapes.

Mach had gone further. He insisted that it is the relative motion of the rotating sphere with respect to the distant masses that was responsible for the observed bulging of the spherical surface. The statement that the "average mass" of the universe gives rise to the inertia of an object has come to be called **Mach's principle**. The question of whether Einstein's final formulation of GR actually incorporates Mach's principle is still being debated. For a recent discussion see, for example, Wilczek (2004), who emphasized that even in Einstein's theory not all coordinate systems are on equal footing.[8] Thus the reader should be aware that there are subtle points with respect to the foundation questions of GR that are still topics in modern theoretical physics research.

1.2.2 Geometry as gravity

Einstein, starting with the EP, made the bold inference that the proper mathematical representation of the gravitational field is a **curved spacetime** (see Chapter 5). As a result, while spacetime has always played a passive role in

[8] This is related to the fact that Einstein's theory is a geometric theory restricted to a metric field, as to be discussed below.

our physics description, it has become dynamic quantity in GR. Recall our experience with electromagnetism; a field theoretical description is a two-step description: the source, i.e. a proton, gives rise to field everywhere, as described by the **field equations** (e.g. the Maxwell's equations); the field then acts locally on the test particle, i.e. an electron, to determine its motion, as dictated by the **equation of motion** (Lorentz force law).

$$\text{source} \longrightarrow \text{field} \longrightarrow \text{test particle}$$

GR as a field theory of gravity with curved spacetime as the gravitational field offers the same two-step description. Its essence is nicely captured in an aphorism (by John A. Wheeler):

> Spacetime tells matter how to move
> Matter tells spacetime how to curve

Since a test particle's motion in a curved space follows "the shortest possible and the straightest possible trajectory" (called the **geodesic curve**), the GR equation of motion is the **geodesic equation** (see Sections 4.2, 5.2, and 12.1). The GR field equation (the **Einstein equation**) tells us how the source of mass/energy can give rise to a curved space by fixing the curvature of the space (Sections 5.3 and 12.2). This is what we mean by saying that "GR is a geometric theory of gravity," or "gravity is the structure of spacetime."

1.2.3 Mathematical language of relativity

Our presentation will be such that the necessary mathematics are introduced as they are needed. Ultimately what is required for the study of GR is Riemannian geometry.

Tensor formalism Tensors are mathematical objects having definite transformation properties under coordinate transformations. The simplest examples are scalars and vector components. The principle of relativity says that physics equations should be covariant under coordinate transformation. To ensure that this principle is **automatically** satisfied, all one needs to do is to write physics equations in terms of tensors. Because each term of the equation transforms in the same way, the equation automatically keeps the same form (its covariant) under coordinate transformations. Let us illustrate this point by the familiar example of $F_i = ma_i$ as a rotational symmetric equation. Because every term of the equation is a vector, under a rotation the same relation $F'_i = ma'_i$ holds in the new coordinate system. The physics is unchanged. We say this physics equation possesses the rotational symmetry. (See Section 2.1.1 for more detail.) In relativity, we shall work with tensors that have definite transformation properties under the ever more general coordinate transformations: the Lorentz transformations and general coordinate transformations (see Chapters 10 and 11). If physics equations are written as tensor equations, then they are automatically relativistic. This is why tensor formalism is needed for the study of relativity.

Our presentation will be done in the coordinate-based component formalism. Although, this is somewhat more cumbersome than the coordinate-independent formulation of differential geometry. This choice is made so that the reader can study the physics of GR without overcoming the hurdle of another layer of abstraction.

Metric description vs. full tensor formulation Mathematically understanding the structure of the Einstein equation is more difficult because it involves the Riemannian curvature tensor. A detailed discussion of the GR field equation and the ways of solving it in several simple situations will be postponed till Part III. In Part I, our presentation will be restricted mainly to the description of the space and time in the form of the metric function, which is a mathematical quantity that (for a given coordinate system used to label the points in the space) describes the shape of the space through length measurements. From the metric function one can deduce the corresponding geodesic equation required for various applications (including the study of cosmology in Part II). We will demonstrate in Part III that the metric functions used in Parts I and II are the solutions of Einstein field equation.

In this introductory chapter, we have emphasized the viewpoint of relativity as the coordinate symmetry. We can ensure that physics equations are covariant under coordinate transformations if they are written as tensor equations. Since the tensor formalism will not be fully explicated until Part III, this also means that the symmetry approach will not be properly developed until later in the book, in Chapters 10–12.

1.2.4 GR is the framework for cosmology

The universe is a huge collection of matter and energy. The study of its structure and evolution, the subject of cosmology, has to be carried out in the framework of GR. The large collection of matter and field means we must deal with strong gravitational effects, and to understand its evolution, the study cannot be carried out in the static field theory. The Newtonian theory for a weak and static gravitational field will not be an adequate framework for modern cosmology. In fact, the very basic description of the universe is now couched in the geometric language of general relativity. A "closed universe" is the one having positive spatial curvature, an "open universe" is negatively curved, etc. Thus for a proper study of cosmology, we must first learn GR.

Review questions

1. What is relativity? What is the principle of special relativity? What is general relativity?

2. What is a symmetry in physics? Explain how the statement that no physical measurement can detect a particular physical feature (e.g. orientation, or the constant velocity of a lab), is a statement about a symmetry in physics. Illustrate your explanation with the examples of rotation symmetry, and the coordinate symmetry of SR.

3. In general terms, what is a tensor? Explain how a physics equation, when written in terms of tensors, automatically displays the relevant coordinate symmetry.

4. What are inertial frames of reference? Answer this in three ways.

5. Equations of Newtonian physics are unchanged when we change the coordinates from one to another inertial frame. What is this coordinate transformation? Equations of electrodynamics are unchanged under another set of coordinate transformations. How are these two sets of transformations related? (Need only to give their names and a qualitative description of their relation.)

6. What is the key difference between the coordinate transformations in special relativity and those in general relativity?

7. What motivated Einstein to pursue the extension of special relativity to general relativity?

8. In the general relativistic theory of gravitation, what is identified as the gravitational field? What is the general relativity field equation? The general relativity equation of motion? (Again, only the names.)

9. How does the concept of space differ in Newtonian physics and in Einsteinian (general) relativistic physics?

2 Special relativity and the flat spacetime

- We follow the historical introduction of special relativity (SR) as the symmetry of Maxwell's theory of electromagnetism.
- Einstein proposed a new kinematics: passage of time is different in different inertial frames. The constancy of the speed of light in every inertial frame implies a new invariant spacetime interval.
- A new geometric description interprets the new invariant interval as the length in the 4D pseudo-Euclidean flat manifold, called Minkowski spacetime.
- Transformations among inertial frames can be interpreted as "rotations" in the 4D spacetime, and the explicit form of Lorentz transformations derived.
- Time-dilation and length contraction are the physics consequence of a spacetime manifold with a metric matrix equal to diag(-1,1,1,1).

In this chapter, a brief discussion of special relativity (SR) is presented. We clarify its conceptual foundation and introduce the geometric formalism in terms of flat spacetime. This prepares us for the study of the larger framework of curved spacetime in general relativity (GR).

2.1 Coordinate symmetries

In Chapter 1 we have already introduced the concept of a symmetry in physics. It is the situation when physics equations, under some transformation, are unchanged in their form (i.e. they are "covariant").[1] Here we shall first review the familiar case of rotational symmetry, in preparation for our discussion of Galilean symmetry of classical mechanics, and Lorentz symmetry of electrodynamics. We shall discuss the distinction between Galilean and Lorentz transformations, first their formal aspects in this section, then their physical basis in Section 2.2. In particular, we first introduce the Lorentz symmetry as some mathematical property of the electrodynamics equation. Only afterwards do we, following Einstein's teachings, discuss the physics as implied by such a coordinate symmetry.

2.1.1 Rotational symmetry

We shall illustrate the statement about symmetry with the familiar example of rotational invariance. To have rotational symmetry means that physics is

[1] Under a transformation, an "invariant" quantity does not change; a "covariant" quantity 'changes in the same way'. Thus, if all terms in an equation are covariant, their relation, hence the equation, is unchanged. The equation is said to be "covariant under the transformation".

unchanged under a rotation of coordinates (NB, not a rotating coordinate). Take the equation of $F_i = ma_i$ $(i = 1, 2, 3)$, which is the familiar $\mathbf{F} = m\mathbf{a}$ equation in the component notation, see Box 2.1. The same equation holds in different coordinate frames which are rotated with respect to each other. Namely, the validity of $F_i = ma_i$ in a system O implies the validity of $F'_i = m'a'_i$ in any other systems O' which are related to O by a rotation. Mass m being a scalar, while a_i and F_i being vector components of the acceleration and force, we have

$$m' = m, \quad a'_i = \sum_j [\mathbf{R}]_{ij}\, a_j, \quad F'_i = \sum_j [\mathbf{R}]_{ij}\, F_j, \tag{2.1}$$

where $[\mathbf{R}]_{ij}$ are the elements of the rotational matrix. (See Box 2.1 for details.) Thus the validity $F'_i - m'a'_i = 0$ follows from $F_i - ma_i = 0$ because the transformation matrix $[\mathbf{R}]$ is the same for each set of vector components (F_i and a_i):

$$F'_i - m'a'_i = \sum_j [\mathbf{R}]_{ij} \left(F_j - ma_j\right) = 0. \tag{2.2}$$

That each term in this physics equation $F_i = ma_i$ transforms in the same way under the rotational transformation is displayed in Fig. 2.1. Under a transformation, the different components of force and acceleration do change values but their relations are not changed as the physics equation keeps the same form. $\mathbf{F} = m\mathbf{a}$ is a vector equation (or, more generally, a tensor equation) as each term of the equation has the same transformation property (as a vector) under rotation. We see that if the physics equation can be written as a vector equation, it automatically respects rotation symmetry.

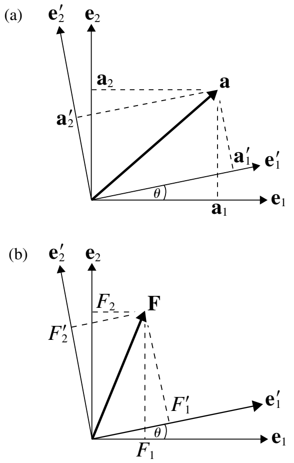

Fig. 2.1 Coordinate change of a vector under rotation. A change of the basis vectors means that components of different vectors, whether acceleration vector as in (a) or force vector as in (b), all transform in the same way, as in (2.1).

Box 2.1 Coordinate transformation in the component notation

For a given coordinate system with basis vectors $\{\mathbf{e}_i\}$, a vector—for example, the position vector $\mathbf{x}$—can be represented by its components x_1, x_2, and x_3, with $\{x_i\}$ being the coefficients of expansion of $\mathbf{x}$ with respect to the basis vectors:

$$\mathbf{x} = \sum_{i=1}^{3} x_i \mathbf{e}_i = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + x_3\mathbf{e}_3. \tag{2.3}$$

With a change of the coordinate system $\{\mathbf{e}_i\} \rightarrow \{\mathbf{e}'_i\}$, the same vector would be represented by another set of components (Fig. 2.1):

$$\mathbf{x} = \sum_{i=1}^{3} x'_i \mathbf{e}'_i. \tag{2.4}$$

For the example of the coordinate transformation being a rotation by an angle of θ around the z-axis, the new position components are related to the original ones by the relation as can be worked out geometrically from Fig. 2.1:

$$\begin{aligned} x'_1 &= \cos\theta x_1 + \sin\theta x_2, \\ x'_2 &= -\sin\theta x_1 + \cos\theta x_2, \\ x'_3 &= x_3. \end{aligned} \tag{2.5}$$

This set of equations can be written compactly as a matrix (the rotation transformation matrix) multiplying the original vector to yield the new position components:

$$\begin{pmatrix} x_1' \\ x_2' \\ x_3' \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}. \tag{2.6}$$

This matrix equation can be expressed in component notation as

$$x_i' = \sum_{j=1}^{3} [\mathbf{R}]_{ij}\, x_j = [\mathbf{R}]_{i1}\, x_1 + [\mathbf{R}]_{i2}\, x_2 + [\mathbf{R}]_{i3}\, x_3, \tag{2.7}$$

where $[\mathbf{R}]_{11} = \cos\theta$ and $[\mathbf{R}]_{12} = \sin\theta$, etc. Such a transformation holds for **all** the vector components. For example, the components of the acceleration vector **a** and force vector **F** transform in the same way:

$$a_i' = \sum_j [\mathbf{R}]_{ij}\, a_j, \quad F_i' = \sum_j [\mathbf{R}]_{ij}\, F_j \tag{2.8}$$

—**with the same rotation matrix** [**R**] as in (2.6). In fact this is the definition of vector components. Namely, they are a set of numbers $\{V_i\}$, which, under a rotation, changes according to the transformation rule given in (2.7) and (2.8):

$$V_i' = \sum_j [\mathbf{R}]_{ij}\, V_j. \tag{2.9}$$

2.1.2 Newtonian physics and Galilean symmetry

One of the most important lessons Galileo and Newton have taught us is that description of the physical world (hence the physics laws) is simplest when using the **inertial frames of reference**. The transformation that allows us to go from one inertial frame O with coordinates x_i to another inertial frame O' with coordinates x_i' is the Galilean transformation: if the relative velocity of the two frames is given to be **v** (a constant) and their relative orientation are specified by three angles α, β, and γ, the new coordinates are related to the old ones by $x_i \longrightarrow x_i' = [\mathbf{R}]_{ij} x_j - v_i t$, where $[\mathbf{R}] = [\mathbf{R}(\alpha, \beta, \gamma)]$ is the rotation matrix. In Newtonian physics, the time coordinate is assumed to be absolute, that is, it is the same in every coordinate frame. In the following we shall be mainly interested in coordinate transformations among inertial frames with the same orientation, $[\mathbf{R}(0,0,0)]_{ij} = \delta_{ij}$ (see Fig. 2.2). Such a transformation is called a (Galilean) **boost**:

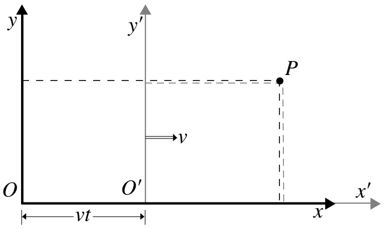

Fig. 2.2 The point P is located at (x, y) in the O system and at (x', y') in the O' system which is moving with velocity v in the x direction. We have $x' = x - vt$, and $y' = y$.

$$\begin{aligned} \mathbf{x} &\longrightarrow \mathbf{x}' = \mathbf{x} - \mathbf{v}t, \\ t &\longrightarrow t' = t. \end{aligned} \tag{2.10}$$

Newtonian relativity says that the physics laws (mechanics) are unchanged under the Galilean transformation (2.10). Physically this implies that no mechanical experiment can detect any intrinsic difference between the two inertial frames. It is easy to check that a physics equation such as $Gm\hat{\mathbf{r}}/r^2 = \mathbf{a}$ does not change its form under the Galilean transformation as the coordinate separation $\mathbf{r} = \mathbf{x}_1 - \mathbf{x}_2$ and the acceleration $\mathbf{a}$ are unchanged under this transformation. In contrast to these invariant quantities, the velocity vector, $\mathbf{u} = d\mathbf{x}/dt$, will change: it obeys the **velocity addition rule**:

$$\mathbf{u} \longrightarrow \mathbf{u}' = \mathbf{u} - \mathbf{v}, \tag{2.11}$$

which is obtained by a differentiation of (2.10).

2.1.3 Electrodynamics and Lorentz symmetry

One can show that Maxwell's equations are not covariant under Galilean transformation. The easiest way to see this is by recalling the fact that the propagation speed of the electromagnetic wave is a constant

$$c = \sqrt{\frac{1}{\mu_0 \epsilon_0}} \tag{2.12}$$

with ϵ_0 and μ_0, the permittivity and permeability of free space, being the constants appearing in the Coulomb's and Ampere's laws. Clearly, c is the same in all inertial frames. This constancy violates the Galilean velocity addition rule of (2.11). Two **alternative** interpretations can be drawn from this apparent violation:

1. Maxwell's equations do not obey the principle of (Newtonian) relativity. By this we mean that Maxwell's equations are valid only in one inertial frame. Hence the relativity principle is not applicable. It was thought that, like all mechanical waves, an electromagnetic wave must have an elastic medium for its propagation. Maxwell's equations were thought to be valid only in the rest frame of the **aether** medium. The constant c was interpreted to be the wave speed in this aether—the frame of **absolute rest**. This was the interpretation accepted by most of the nineteen century physicists.
2. Maxwell's equations do obey the principle of relativity but the relation among inertial frames are not correctly given by the Galilean transformation. Hence the velocity addition rule of (2.11) is invalid; the correct relation should be such that c can be the same in every inertial frame. The modification of velocity addition rule must necessarily bring about a change of the time coordinate $t' \neq t$.

The second interpretation turned out to be correct. The measurement made by Michelson and Morley showed that the speed of light is the same in different moving frames. It had been discovered by Poincaré, independent of Einstein's 1905 work, that Maxwell's equations were covariant under a new boost transformation, "the **Lorentz transformation**" (see Box 2.2). Namely, Maxwell's equations keep the same form if one makes the formal change, including the time variable, from (t, x, y, z) to (t', x', y', z'), representing the coordinates of two frames moving with a relative velocity $\mathbf{v} = v\hat{\mathbf{x}}$:

$$x' = \gamma(x - vt), \quad y' = y, \quad z' = z, \quad t' = \gamma\left(t - \frac{v}{c^2}x\right), \tag{2.13}$$

where the parameter γ depends on the relative speed of the two reference frames as

$$\gamma = \frac{1}{\sqrt{1 - \beta^2}}, \qquad \beta = \frac{v}{c}. \tag{2.14}$$

We have in general $\beta \leq 1$ and $\gamma \geq 1$. We note that the spatial transformation is just the Galilean transformation (2.10) multiplied by the γ factor and is thus reduced to (2.10) in the low velocity limit of $\beta \to 0$, hence $\gamma \to 1$.

Box 2.2 Maxwell's equations and Lorentz transformation

An electric charge at rest gives rise to an electric, but not a magnetic, field. However, the same situation when seen by a moving observer is a charge in motion, which produces both electric and magnetic fields. This shows that the electric and magnetic fields will change into each other in moving coordinates. When we say Maxwell's equation is covariant under the Lorentz transformation we must also specify the Lorentz transformation properties of the fields **E** and **B** as well as the current and charge densities, **j** and ρ. Namely, under Lorentz transformation, not only the space and time coordinates will change, but also the electromagnetic fields and source charge and currents. However, the relation among these changed quantities remain the same, as those given by the Maxwell's equations as in the original frame of reference.

The transformation formulae for these electromagnetic quantities are somewhat simpler when written in the Heaviside–Lorentz system of units,[2] in which the measured parameter is taken, instead of (ϵ_0, μ_0), to be c, the velocity of EM wave. In this system, the Lorentz force law reads

$$\mathbf{F} = q\left(\mathbf{E} + \frac{1}{c}\mathbf{v} \times \mathbf{B}\right), \tag{2.15}$$

while the Maxwell equations take the form

$$\nabla \cdot \mathbf{B} = 0, \quad \nabla \times \mathbf{E} + \frac{1}{c}\frac{\partial \mathbf{B}}{\partial t} = 0, \tag{2.16}$$

$$\nabla \cdot \mathbf{E} = \rho, \quad \nabla \times \mathbf{B} - \frac{1}{c}\frac{\partial \mathbf{E}}{\partial t} = \frac{\mathbf{j}}{c}. \tag{2.17}$$

In this unit system, the Lorentz transformation properties of the electromagnetic fields are given by

$$\begin{aligned} E_1' &= E_1, \quad E_2' = \gamma(E_2 - \beta B_3), \quad E_3' = \gamma(E_3 + \beta B_2), \\ B_1' &= B_1, \quad B_2' = \gamma(B_2 + \beta E_3), \quad B_3' = \gamma(B_3 - \beta E_2) \end{aligned} \tag{2.18}$$

and those of the charge and current densities are given by

$$j_1' = \gamma\,(j_1 - v\rho)\,, \quad j_2' = j_2, \quad j_3' = j_3, \quad \rho' = \gamma\left(\rho - \frac{v}{c^2}j_1\right). \tag{2.19}$$

Using these transformation rules, as well as those for the space and time coordinates (2.13), we can check in a straightforward manner (as in Problem 2.4) that equations of electromagnetism are unchanged in their form (i.e. they are covariant) under Lorentz transformation.

2.1.4 Velocity addition rule amended

Maxwell's equations respect the Lorentz symmetry, and they must be compatible with the physical phenomenon of the electromagnetic wave propagating with the same velocity c in all the moving frames. The Lorentz transformation

[2] Conversion table from mks unit system to that of Heaviside–Lorentz:

mks		Heaviside–Lorentz
$\sqrt{\epsilon_0}\,\mathbf{E}$	$\longrightarrow$	$\mathbf{E}$
$\sqrt{1/\mu_0}\,\mathbf{B}$	$\longrightarrow$	$\mathbf{B}$
$\sqrt{1/\epsilon_0}(\rho, \mathbf{j})$	$\longrightarrow$	$(\rho, \mathbf{j})$

must imply a new velocity addition rule which allows for a constant c in every inertial frame. Writing (2.13) in differential form

$$dx' = \gamma(dx - vdt), \quad dy' = dy, \quad dz' = dz, \quad dt' = \gamma\left(dt - \frac{v}{c^2}dx\right), \tag{2.20}$$

we obtain the velocity transformation rule by simply constructing the appropriate quotients:

$$u'_x = \frac{dx'}{dt'} = \frac{dx - vdt}{dt - (v/c^2)dx} = \frac{u_x - v}{1 - (vu_x/c^2)}, \tag{2.21}$$

$$u'_y = \frac{dy'}{dt'} = \frac{dy}{\gamma(dt - (v/c^2)dx)} = \frac{u_y}{\gamma(1 - (vu_x/c^2))}, \tag{2.22}$$

$$u'_z = \frac{dz'}{dt'} = \frac{u_z}{\gamma(1 - (vu_x/c^2))}. \tag{2.23}$$

For the special case of two frames moving with a relative velocity $\mathbf{v} = v\hat{x}$ parallel to the velocity under study: $u_x = u$ and $u_y = u_z = 0$, we have

$$u' = \frac{u - v}{1 - ((uv)/c^2)}, \tag{2.24}$$

while the y and z components remain unchanged ($u'_y = u'_z = 0$). Namely, it is just the familiar velocity addition (2.11), but with the right-hand side (RHS) divided by an extra factor of $(1 - uv/c^2)$. It is easy to check that an input of $u = c$ leads to an output of $u' = c$—thus the constancy of the light velocity in every inertial frame of reference.

The Michelson–Morley experiment confirmed the notion that speed of light c is the same in different inertial frames. Namely, the Galilean velocity addition rule Eq. (2.11) is not obeyed. But historically, because Einstein had already been convinced of the physical validity of a constant c, this experimental result per se did not play a significant role in Einstein's thinking when he developed the theory of SR.

2.2 The new kinematics of space and time

The covariance under Lorentz transformation, that is, the coordinate independent nature, of electromagnetism equations was independently discovered by Poincaré. But it was Einstein who had first emphasized the physical basis of a new kinematics that was required to fully implement the new symmetry—in particular the necessity of having different time coordinates in different inertial frames when the speed of signal transmission was not infinite. He had emphasized that the definition of time was ultimately based on the notion of **simultaneity** because the requirement of clock synchronization, etc., but simultaneity ($\Delta t = 0$, actually any definite time interval Δt) is not absolute, when the speed of signal transmission is finite. Namely, a time interval measured by one inertial observer will differ from that by another who is in relative motion with respect to the first observer. **Simultaneity is also a relative concept.**

A coordinate system is a reference system with a coordinate grid (to determine the position) and a set of clocks (to determine the time of an event). We require all the clocks to be synchronized (say, against the master clock located at the origin). The synchronization of a clock, located at a distance r from the origin, can be accomplished by sending out light flashes from the master clock at $t = 0$. When the clock receives the light signal, it should be set at $t = r/c$. Equivalently, synchronization of any two clocks can be checked by sending out light flashes from these two clocks at a given time. If the two flashes arrive at their midpoint at the same time, they are synchronized.

2.2.1 Relativity of spatial equilocality

To describe a certain quantity as being relative means that it is not invariant under coordinate transformations. In this section, we shall consider the various invariants (and noninvariants) under different types of coordinate transformations.

Two events happening at the same spatial location are termed to be "equilocal." If two events (x, t_1) and (x, t_2) do not take place at the same time, $\Delta t = t_2 - t_1 \neq 0$, equilocality for these two events is already a relative notion even under Galilean transformation (2.10),

$$\Delta x' = v\Delta t \neq 0, \qquad \text{even though } \Delta x = 0. \tag{2.25}$$

It is useful to have a specific illustration: a light bulb at a fixed position on a moving train emits two flashes of light. To an observer on the train these two events are spatially equilocal but not simultaneous. Clearly this equilocality is a relative concept, because, to an observer standing on the rail platform as the train passes by, they appear as two flashes emitted at two different locations. See Fig. 2.3.

2.2.2 Relativity of simultaneity—the new kinematics

Einstein pointed out that, in reality where the signal transmission could not be carried out at infinite speed, simultaneity of two events would be a relative concept: two events, observed by one observer to be simultaneous, would be seen by another observer in relative motion to occur at different times.

First we need a commonly agreed-upon definition of simultaneity. For example, we can mark off the midpoint between two locations. Two events that take place at these two locations are said to be simultaneous if they are "seen" by the observer at the midpoint to take place at the same time. The operation

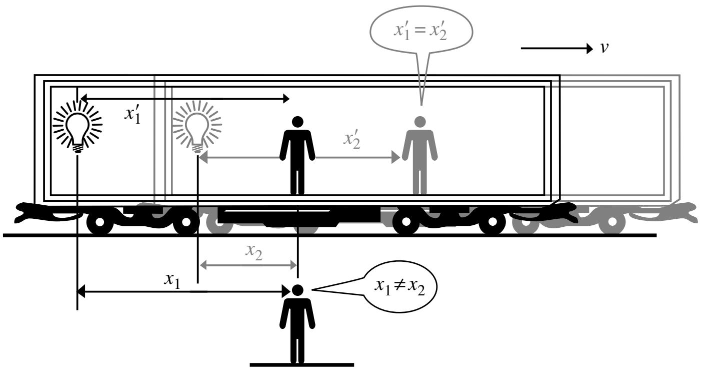

Fig. 2.3 Spatial congruity of two events is relative if they take place at different times. A light bulb at a fixed position on a moving train flashes twice. To the observer on the train, these two events (x_1', t_1') and (x_2', t_2') are spatially congruous $x_1' = x_2'$; but to another observer standing on the rail platform, these two events (x_1, t_1) and (x_2, t_2) take place at two different locations $x_1 \neq x_2$.

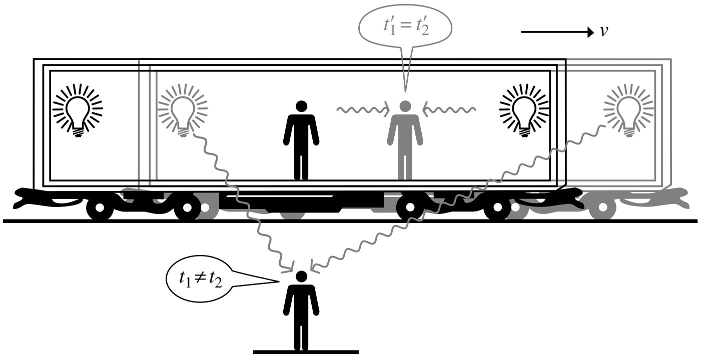

Fig. 2.4 Simultaneity is relative when light is not transmitted instantaneously. Two events (x'_1, t'_1) and (x'_2, t'_2) corresponding to lights flashed at opposite ends of a moving train are seen as simultaneous $t'_1 = t'_2$ by an observer on the train (e.g. with the observer receiving the signal simultaneously when standing at the midpoint). But to another observer standing on the rail platform, these two events (x_1, t_1) and (x_2, t_2) are not simultaneous, $t_1 \neq t_2$, because the light signals reach her at different times.

of "seeing" these two events involves receiving light signals from these two events. Apply this operational definition of simultaneity to the following case. Two light bulbs are located certain distance apart $\Delta x'$. If an observer standing midway receives light signals from these two bulbs at the same time, this observer will regard the emissions from these two light bulbs as simultaneous events. Namely, the observer would deduce that these two events of light emission took place at two equal intervals ago: $t'_1 = t'_2 = \Delta x'/2c$.

We now illustrate the relativity of simultaneity for two observers in relative motion (Fig. 2.4). Let these two light bulbs be located at the two ends of a rail car. One observer is at the midpoint on the moving car, another observer at midpoint on the rail platform. (One can pre-arrange triggers on the rail so that the bulbs emit their light signals when the rail car's middle just line up with the platform observer. Namely, the lights originate at equal distance from the observer). As a result of the moving car and the finite light speed, the emissions, seen by the rail car observer to be simultaneous, will no longer be seen by the platform observer to be simultaneous. When the light pulses arrive at the observer, this would no longer be the midpoint: one bulb would have moved further away and the other closer. It would then take different amounts of time to cover these two different distances, resulting in different arrival times at the platform observer. To this observer these two emission events are not simultaneous.

Let us calculate the deviation from simultaneity as seen by the platform observer. For an observer, the time interval it takes light to travel the distance from the bulb to the observer is their distance separation (at the time of light arrival) divided by the speed of light. The initial separation between the two bulbs being the rail-car length L_p as seen by the platform observer,[3] the distance at the arriving time between the "approaching bulb" and the observer is $\frac{1}{2}L_\mathrm{p} - vt$, and the distance to the "receding bulb" is $\frac{1}{2}L_\mathrm{p} + vt$. Divided by c, they give rise to two different time intervals t_1 and t_2. Their difference is the amount of nonsimultaneity:

$$t_2 - t_1 = \frac{L_\mathrm{p}}{2c}\left(\frac{1}{1-\beta} - \frac{1}{1+\beta}\right) = \gamma^2 \frac{\beta}{c} L_\mathrm{p}, \tag{2.26}$$

where we have used the expression of β and γ of (2.14). Namely two events, seen in one frame to be simultaneous $\Delta t' = 0$, are observed by a moving

[3] We simplify the kinematics to an 1D problem by assuming negligibly small transverse lengths.

observer to take place at two instances apart, by

$$\Delta t = \gamma^2 \frac{\beta}{c} L_{\mathrm{p}}. \tag{2.27}$$

Further calculations relating to the issue of simultaneity can be found in Section 2.3.4, as well as in Problem 2.15.

The reason why one reached the erroneous conclusion in Newtonian physics (that the rail platform observer also sees these as two simultaneous events) is related to the fact that the train speed is extremely small compared to the speed of light signal propagation, $v \ll c$. Namely, the nonsimultaneity in the rail platform frame is so small, of the order of v/c, as to be unobservable. The true transformation rule can in the low-speed limit be approximated by taking the limit of $v/c \to 0$ (namely, $c \to \infty$). This of course reduces the transformation (2.13) to the Galilean form (2.10).

2.2.3 The invariant space–time interval

Now if $\Delta \mathbf{x}$ and Δt are no longer absolute to different observers, is there any invariant left? It turns out that there is still one invariant—a certain combination of $\Delta \mathbf{x}$ and Δt remains to be absolute even though space and time measurements are all relative.

To find this new invariant, we first need to state the basic postulates of SR:

> **Principle of relativity.** Physics laws have the same form in every inertial frame of reference. No physical measurement can reveal the absolute motion of an inertial frame of reference.
>
> **Constancy of the light speed.** This second postulate is certainly consistent with the first one. The constancy of light speed is a feature of electrodynamics and the principle of relativity would lead us to expect it to hold in every frame.

We shall show that the following space–time interval is absolute, that is, it has the same value in every inertial frame (Landau and Lifshitz, 1975).

$$\Delta s^2 = \Delta x^2 + \Delta y^2 + \Delta z^2 - c^2 \Delta t^2, \tag{2.28}$$

where $\Delta x = x_2 - x_1$, etc. (Table 2.1). Ultimately this invariance comes about because of the constancy of c in every reference frame: Δs is absolute because c is absolute.

First consider the **special case** when the two events $(\mathbf{x}_1, t_1)$ to $(\mathbf{x}_2, t_2)$ are connected by a light signal. The interval Δs^2 must vanish because in this case $(\Delta x^2 + \Delta y^2 + \Delta z^2)/\Delta t^2 = c^2$. When observed in another frame O', this interval also has a vanishing value $\Delta s'^2 = 0$, because the velocity of light remains the same in the new frame O'. From this, we conclude that **any** interval connecting two events (not necessarily by a light signal) when viewed in two different coordinates must always be proportional to each other because, if Δs^2 vanishes, so must $\Delta s'^2$:

$$\Delta s'^2 = F \Delta s^2, \tag{2.29}$$

where F is the proportional factor, and it can in principle depend on the coordinates and the relative velocity of these two frames: $F = F(\mathbf{x}, t, \mathbf{v})$. However, the requirement of space and time being homogeneous (i.e. there is no privileged point in space and in time) implies that there cannot be any dependence of **x** and t. That space is isotropic means that the proportional factor cannot depend on the direction of their relative velocity **v**. Thus we can at most have it to be dependent on the magnitude of the relative velocity, $F = F(v)$. We are now ready to show that, in fact, $F(v) = 1$.

Besides the system O', which is moving with velocity of **v** with respect to system O, let us consider yet another inertial system O'' which is moving with a relative velocity of $-\mathbf{v}$ with respect to the O' system.

$$O \xrightarrow{\mathbf{v}} O' \xrightarrow{-\mathbf{v}} O''. \tag{2.30}$$

From the above consideration, and applying (2.29) to these frames:

$$\begin{aligned}\Delta s'^2 &= F(v)\Delta s^2,\\ \Delta s''^2 &= F(v)\Delta s'^2 = [F(v)]^2\Delta s^2.\end{aligned} \tag{2.31}$$

However, it is clear that the O'' system is in fact just the O system. This requires $[F(v)]^2 = 1$. The solution $F(v) = -1$ being nonsensical, we conclude that this interval Δs is indeed an invariant: $\Delta s'' = \Delta s' = \Delta s$. Namely every inertial observer, who always sees the same light velocity, would obtain the same value for this particular combination of space and time interval.

That the space–time combination $s^2 = x^2 + y^2 + z^2 - c^2t^2$ is invariant under Lorentz transformation can be checked by using the explicit form of the transformation rule as given in Eq. (2.13).

Proper time This interval Δs has the physical significance of being directly related to the time interval in the rest frame of the particle: rest frame means there is no spatial displacement $\Delta\mathbf{x} = 0$,

$$\Delta s^2 = -c^2\Delta\tau^2. \tag{2.32}$$

The rest-frame time coordinate τ is called the **proper time**. Since there is only one rest-frame, its time interval must be unique—all observers should agree on its value. This is the physical basis for the invariance of this quantity.

New kinematics and dynamics In Section 2.3 we shall present the new kinematics in which the invariance of the space–time interval Δs plays a key role. The new kinematics is the setting for the coordinate symmetry showing that physics is unchanged under coordinate transformations that have an invariant Δs. Such transformations, the Lorentz transformations, can be thought as "rotations" in the 4D space of three spatial, and one time, coordinates, with a length given by Δs. Maxwell's electrodynamics already has this new relativity

Table 2.1 Intervals that are invariant (marked by ✓) under the respective transformations vs. those that are not

Intervals		Galilean transformation	Lorentz transformation
Δt	for $\Delta x \neq 0$	✓	×
Δx	for $\Delta t \neq 0$	×	×
$\Delta x^2 - c^2\Delta t^2$			✓

symmetry, but Newton's laws of mechanics do not. They will have to be generalized so as to be compatible with this coordinate symmetry. However, this discussion of the relativistic dynamics will be postponed till Chapter 10, when we present tensor formalism of the 4D spacetime.

2.3 Geometric formulation of SR

Maxwell's equations for electrodynamics are not compatible with the principle of Newtonian relativity. Most notably, the constancy of electromagnetic wave velocity in every inertial frame violates the familiar velocity addition rule of (2.11). Consequently it is difficult to formulate a consistent electrodynamic theory for a moving observer. Einstein's resolution of these difficulties was stated in an all-embracing new kinematics.[4] In other words, an understanding of the physics behind the Lorentz covariance, as first discovered in Maxwell's equations, would involve a revision of our basic notions of space and time. This would have fundamental implications for all aspects of physics, far and beyond electromagnetism.

The new kinematics can be expressed elegantly in a geometric formalism of 4D spacetime as first formulated by Herman Minkowski. The following are the opening words of an address he delivered at the 1908 Assembly of German National Scientists and Physicians held in Cologne.

> The views of space and time which I wish to lay before you have sprung from the soil of experimental physics, and therein lies their strength. They are radical. Henceforth space by itself, and time by itself, are doomed to fade away into mere shadows, and only a kind of union of the two will preserve an independent reality.

Special relativity emphasizes the symmetry between space and time. But spatial length and time interval have different measurement units. One way to see the significance of light speed c is that it is the **conversion factor** connecting the space and time coordinates. Thus the dimensionful number c is of fundamental importance to relativity, because without it we would not be able to discuss the symmetry of physics with respect to transformations between space and time.

2.3.1 General coordinates and the metric tensor

We will interpret the new spacetime invariant in (2.28) as the expression of a length in a 4D space with ct being the fourth coordinate. In a 4D Euclidean space with Cartesian coordinates (w, x, y, z), the invariant length is given as $s^2 = w^2 + x^2 + y^2 + z^2$. One the other hand, what we have in SR is $s^2 = -c^2t^2 + x^2 + y^2 + z^2$. The minus sign in front of the c^2t^2 term means that if we choose to think ct being the fourth dimension, we must work with a pseudo-Euclidean space, and consider coordinates different from the familiar Cartesian coordinates. In this section, we shall introduce the topic of generalized coordinates and distance measurements (via the metric). The same formalism is applicable to coordinates in a warped space, which we will need to use in later discussions.

[4] The famous 1905 paper by Einstein had the title, "On the electrodynamics of moving bodies."

Basis vectors define the metric

To set up a coordinate system for an n-dimensional space means to chose a set of basis vectors $\{\mathbf{e}_i\}$ where $i = 1, 2, \ldots, n$. In general this is not an orthonormal set $\mathbf{e}_i \cdot \mathbf{e}_j \neq \delta_{ij}$, where δ_{ij} is the Kronecker delta: it equals 1 when $i = j$ and 0 when $i \neq j$. Nevertheless we can represent it as a symmetric matrix, called the **metric**, or the **metric tensor**:

$$\mathbf{e}_i \cdot \mathbf{e}_j \equiv g_{ij}, \tag{2.33}$$

which can be viewed as the elements of a matrix

$$[\mathbf{g}] = \begin{pmatrix} g_{11} & g_{12} & \cdots \\ g_{21} & g_{22} & \cdots \\ \vdots & \vdots & \end{pmatrix} = \begin{pmatrix} \mathbf{e}_1 \cdot \mathbf{e}_1 & \mathbf{e}_1 \cdot \mathbf{e}_2 & \cdots \\ \mathbf{e}_2 \cdot \mathbf{e}_1 & \mathbf{e}_2 \cdot \mathbf{e}_2 & \cdots \\ \vdots & \vdots & \end{pmatrix}. \tag{2.34}$$

Thus the diagonal elements are the (squared) length of the basis vectors: $|e_1|^2$, $|e_2|^2$, etc., while the off-diagonal elements represent their deviations from orthogonality. Namely, any set of mutually perpendicular bases would be represented by a diagonal metric matrix, even though the bases may not have unit lengths.

Given the definition (2.33), it is clear that metric for a curved space must be position-dependent $g_{ij}(x)$ as in such a space the bases vectors $\{\mathbf{e}_i\}$ must necessarily change from point to point. This means that a flat space is the one in which it is possible to find a coordinate system so that the metric is position independent, that is, all elements of the metric matrix for a flat space are constants. For an Euclidean space, we can have the Cartesian coordinates with a set of orthonormal bases $\mathbf{e}_i \cdot \mathbf{e}_j = \delta_{ij}$. Namely, the metric is simply given by the identity matrix, $[\mathbf{g}] = \mathbf{1}$.

We can expand any vector in terms of the basis vectors

$$\mathbf{V} = \sum_i V^i \mathbf{e}_i, \tag{2.35}$$

where the coefficients of expansion $\{V^i\}$ are labeled with superscript indices.[5] Consider the scalar product of two vectors

$$\begin{aligned} \mathbf{V} \cdot \mathbf{U} &= \left(\sum_i V^i \mathbf{e}_i\right) \cdot \left(\sum_j U^j \mathbf{e}_j\right) \\ &= \sum_{i,j} \mathbf{e}_i \cdot \mathbf{e}_j V^i U^j = \sum_{i,j} g_{ij} V^i U^j. \end{aligned} \tag{2.36}$$

Making it more explicit

$$\mathbf{V} \cdot \mathbf{U} = \left(V^1, V^2, \ldots\right) \begin{pmatrix} g_{11} & g_{12} & \cdots \\ g_{21} & g_{22} & \cdots \\ \vdots & \vdots & \end{pmatrix} \begin{pmatrix} U^1 \\ U^2 \\ \vdots \end{pmatrix}. \tag{2.37}$$

The metric is needed to relate the scalar product to the vector components. For the case $\mathbf{V} = \mathbf{U}$, the above equation is an expression for the (squared) length of the vector. Thus the metric relates the length to the vector components. In fact, a common practice is to define the metric through this relation between the length and coordinates, cf. (2.48).

[5] Such a convention is adopted here in preparation for our discussion of tensors in Chapters 10 and 11, where upper-indexed components are identified as the contravariant components of a vector (or tensor), while the lower-indexed ones are the covariant components.

The summation of an upper index with a lower index is called a “contraction.” At this point we also introduce the “Einstein summation convention”—we omit the display of summation signs, whenever there is a pair of repeated (upper and lower) indices: it is understood that they are being summed over. For example

$$g_{ij}V^iU^j \equiv \sum_{i,j} g_{ij}V^iU^j. \tag{2.38}$$

Coordinate transformations

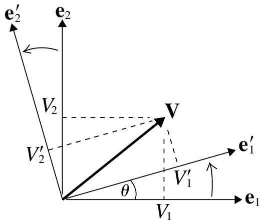

Fig. 2.5 Changing a system $(\mathbf{e}_1, \mathbf{e}_2)$ to another $(\mathbf{e}_1', \mathbf{e}_2')$ in a coordinate transformation.

Let us now consider a coordinate transformation $\{\mathbf{e}_i\} \rightarrow \{\mathbf{e}_i'\}$. Namely, the basis vectors change while the vector quantity, say, $\mathbf{V}$ is unchanged, see Fig. 2.5. On the other hand the vector components $\{V^i\}$, being the projections onto the coordinate bases, would change $\{V^i\} \rightarrow \{V'^i\}$ along with the coordinates. They are related by the transformation matrix elements $[\mathbf{R}]^i_j$ so that

$$V'^i = [\mathbf{R}]^i_{\ j}V^j, \quad V^i = [\bar{\mathbf{R}}]^i_{\ j}V'^j, \tag{2.39}$$

where $[\bar{\mathbf{R}}]$ is the inverse transformation $[\bar{\mathbf{R}}] = [\mathbf{R}]^{-1}$ Rotation transformation as an example has already been discussed in (2.8). NB we have used Einstein’s convention so that the summation of “dummy indices” j is understood. A scalar product being an invariant, we must have, using (2.36),

$$g_{ij}V^iU^j = g'_{kl}V'^kU'^l. \tag{2.40}$$

(Namely, we restrict ourselves to length preserving transformations.) Substitute in the transformation rule of (2.39) on the left-hand side (LHS), we have[6]

$$g_{ij}[\bar{\mathbf{R}}]^i_{\ k}[\bar{\mathbf{R}}]^j_{\ l}V'^kU'^l = g'_{kl}V'^kU'^l. \tag{2.41}$$

Since the vector components V'^k and U'^l are completely arbitrary, in order for the above equality to hold for all possible values of V'^k and U'^l, their coefficients must equal:

$$g_{ij}[\bar{\mathbf{R}}]^i_{\ k}[\bar{\mathbf{R}}]^j_{\ l} = g'_{kl}. \tag{2.42}$$

A multiplication of two matrices involves a summation of the column index of the first matrix and the row index of the second. Namely, the summed index pair must stand next to each other. In the above expression (2.42), while the pair of j indices satisfy this condition and there is the multiplication of the metric matrix $[\mathbf{g}]$ and the transformation matrix $[\bar{\mathbf{R}}]$

$$g_{ij}[\bar{\mathbf{R}}]^j_{\ l} = ([\mathbf{g}][\bar{\mathbf{R}}])_{il}, \tag{2.43}$$

the i-pair indices are out of order. For this to represent a matrix multiplication, we must flip the order of the two indices (i, k)—interchange the row and column of the $[\bar{\mathbf{R}}]$ matrix:

$$g_{ij}[\bar{\mathbf{R}}]^i_{\ k} = [\bar{\mathbf{R}}]^i_{\ k}g_{ij} = \left[\bar{\mathbf{R}}^\top\right]_k^{\ i} g_{ij}, \tag{2.44}$$

where $[\bar{\mathbf{R}}^\top]$ is the transpose of $[\bar{\mathbf{R}}]$. In this way, (2.42) may be written as a matrix equation,

$$[\bar{\mathbf{R}}^\top][\mathbf{g}][\bar{\mathbf{R}}] = [\mathbf{g}'], \tag{2.45}$$

representing the condition for the invariance of scalar products (such as $\mathbf{V} \cdot \mathbf{U}$). Namely, this equation shows how the metric tensor $[\mathbf{g}]$ must change under the coordinate transformation (2.39) that keeps the length of a vector invariant. That it is the inverse transformation $[\bar{\mathbf{R}}]$ that should appear here will be further explained in the tensor chapters (10 and 11). For the Euclidean space with

[6] NB while matrix multiplication is in general noncommutative $[\mathbf{A}][\mathbf{B}] \neq [\mathbf{B}][\mathbf{A}]$, matrix elements, such as $[\mathbf{A}]^j_i$ and $[\mathbf{B}]^k_{\ j}$, being ordinary numbers, can be written in whatever order we wish: $[\mathbf{A}]^j_i[\mathbf{B}]^k_{\ j} = [\mathbf{B}]^k_{\ j}[\mathbf{A}]^j_i$.

Cartesian coordinates we have orthonormal bases so that $[\mathbf{g}'] = [\mathbf{g}] = \mathbf{1}$. (2.45) is reduced to the familiar orthogonality condition $[\bar{\mathbf{R}}^\top][\bar{\mathbf{R}}] = [\mathbf{R}^\top][\mathbf{R}] = \mathbf{1}$. Namely, the length preserving transformation in a Euclidean flat space must be an orthogonal transformation—a rotation. See Section 2.3.2. In a generalized flat space, such as the Minkowski space of SR, while $[\mathbf{g}] \neq \mathbf{1}$, we still have an invariant metric $[\mathbf{g}'] = [\mathbf{g}]$. The generalized orthogonality condition of a length preserving transformation becomes

$$[\mathbf{R}][\mathbf{g}][\mathbf{R}^\top] = [\mathbf{g}], \tag{2.46}$$

which is equivalent to $[\bar{\mathbf{R}}][\mathbf{g}][\bar{\mathbf{R}}^\top] = [\mathbf{g}]$.

Minkowski space and its metric

That measurement results for time, as well as space, are coordinate-frame-dependent means that time can be treated on the same footing as space coordinates. The unification of space and time can be made explicit when space and time coordinates appear in the same vector. We shall reiterate the above discussion, now explicitly for the Minkowski spacetime. This four-dimensional space has coordinates $\{x^\mu\}$ where the index $\mu = 0, 1, 2, 3$. Namely,

$$x^\mu = (x^0,\ x^1,\ x^2,\ x^3) = (ct, x, y, z). \tag{2.47}$$

For the scalar product $\mathbf{V} \cdot \mathbf{U} = g_{ij} V^i U^j$ we consider in particular the infinitesimal interval $ds^2 = d\mathbf{x} \cdot d\mathbf{x} = -c^2 dt^2 + dx^2 + dy^2 + dz^2$, where $d\mathbf{x}$ is a vector in 4D Minkowski space, i.e. a 4-vector. We can interpret the Lorentz transformation geometrically as a "rotation" in the Minkowski space that preserves the length $ds'^2 = ds^2$,

$$\begin{aligned} ds^2 &= (dx^0, dx^1, dx^2, dx^3) \begin{pmatrix} -1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{pmatrix} \begin{pmatrix} dx^0 \\ dx^1 \\ dx^2 \\ dx^3 \end{pmatrix} \\ &\equiv \eta_{\mu\nu} dx^\mu dx^\nu, \end{aligned} \tag{2.48}$$

where we have used the Einstein summation convention in writing the last line. Thus the Minkowski space has a metric

$$g_{\mu\nu} = \mathrm{diag}(-1, 1, 1, 1) \equiv \eta_{\mu\nu}. \tag{2.49}$$

Because the metric is a constant, we say Minkowski spacetime is flat space.[7] It differs from the familiar Euclidean space only by having a negative value for the metric component $\eta_{00} = -1$. As we shall discuss, spacetime manifold is warped in the presence of matter and energy. In fact, in Einstein's general theory of relativity, the curved spacetime is the gravitational field and the metric $g_{\mu\nu}(x)$ is necessarily position-dependent. The pseudo-Euclidean flat spacetime is obtained only in the **absence** of gravity. This is the limit of SR.

Lorentz coordinate transformation $d\mathbf{x}' = [\mathbf{L}]d\mathbf{x}$ may be written in the matrix form as (2.39):

$$dx'^\mu = [\mathbf{L}]^\mu_\nu dx^\nu, \tag{2.50}$$

where $d\mathbf{x}$ is a 4-vector and $[\mathbf{L}]$ is the 4×4 Lorentz transformation matrix. It may be regarded as a rotation in Minkowski space. Elements of $[\mathbf{L}]$ can be fixed

[7] This is a sufficient but not necessary condition for a flat space. As we shall discuss further in Chapter 4 (see specially Box 4.1), only in a flat space we can find a coordinate system with a position independent metric. Consider the example of a flat plane, where the metric is constant in the Cartesian coordinates but not so in the polar coordinate system.

by the length-invariance condition (2.46), now written as

$$[\mathbf{L}][\boldsymbol{\eta}][\mathbf{L}]^\top = [\boldsymbol{\eta}]. \tag{2.51}$$

where $[\boldsymbol{\eta}]$ is the Minkowski metric matrix of (2.49).

2.3.2 Derivation of Lorentz transformation

In the following we shall demonstrate how the pseudo-Euclidean metric of Eq. (2.49) determine, through Eq. (2.51), the form of the length preserving coordinate transformations in the Minkowski spacetime.

The rotation transformation

We start with a rotation around the z-axis by an angle θ that leaves the (z, t) coordinates unchanged. Suppressing such unchanged coordinates, Eq. (2.50) is represented by an effective 2×2 rotation matrix:

$$\begin{pmatrix} dx' \\ dy' \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} dx \\ dy \end{pmatrix}, \tag{2.52}$$

with Eq. (2.51) written out as

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & c \\ b & d \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{2.53}$$

The diagonal conditions of $a^2 + b^2 = c^2 + d^2 = 1$ can be solved by the parametrization of $a = \cos\phi$, $b = \sin\phi$ and $c = \sin\phi', d = \cos\phi'$; while the off-diagonal condition of $ac + bd = \sin(\phi + \phi') = 0$ implies $\phi = -\phi'$. In terms of the actual rotation angle θ, we have the identification $\phi = -\theta$.

$$\begin{pmatrix} dx' \\ dy' \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} dx \\ dy \end{pmatrix}. \tag{2.54}$$

Or, $d\mathbf{x}' = [\mathbf{R}(\theta)]d\mathbf{x}$ and the rotational matrix $[\mathbf{R}(\theta)]$ can be deduced from the length preserving condition (2.51).

The boost transformation

We now consider the relation between two inertial frames connected by a boost (with velocity v) in the $+x$ direction. This coordinate transformation matrix can be similarly fixed by (2.51). Since the (y, z) coordinates are not affected, we again have effectively a two-dimensional problem. Equation (2.50) takes on the form:

$$\begin{pmatrix} cdt' \\ dx' \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} cdt \\ dx \end{pmatrix} \tag{2.55}$$

and the length invariant condition of (2.51) is written out now as

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & c \\ b & d \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{2.56}$$

The conditions of $a^2 - b^2 = -c^2 + d^2 = 1$ can be solved by the parametrization of $a = \cosh\psi, b = \sinh\psi$ and $c = \sinh\psi', d = \cosh\psi'$; while the off-diagonal condition of $-ac + bd = -\cosh\psi\sinh\psi' + \sinh\psi\cosh\psi' = \sinh(\psi - \psi') = 0$ yields $\psi = \psi'$. Thus a Lorentz boost transformation has the matrix form of

$$[\mathbf{L}(\psi)] \equiv \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \cosh\psi & \sinh\psi \\ \sinh\psi & \cosh\psi \end{pmatrix}. \tag{2.57}$$

To relate the parameter ψ to the boost velocity v, we note that the coordinate origin $x' = 0$ of the O' system ($x' = ct\sinh\psi + x\cosh\psi = 0$) moves with velocity $v = x/t$ along the x-axis of the O system.

$$\frac{x}{t} = -c\frac{\sinh\psi}{\cosh\psi} = v \quad \text{or} \quad \frac{\sinh\psi}{\cosh\psi} = -\beta. \tag{2.58}$$

Rewriting the identity $\cosh^2\psi - \sinh^2\psi = 1$ as

$$\cosh\psi\sqrt{1 - \left(\frac{\sinh^2\psi}{\cosh^2\psi}\right)} = 1,$$

we find

$$\cosh\psi = \gamma \quad \text{and} \quad \sinh\psi = -\beta\cosh\psi = -\beta\gamma, \tag{2.59}$$

where β and γ are defined in (2.14). The coordinate transformation (2.55) is found to be

$$\begin{pmatrix} cdt' \\ dx' \end{pmatrix} = \gamma\begin{pmatrix} 1 & -\beta \\ -\beta & 1 \end{pmatrix}\begin{pmatrix} cdt \\ dx \end{pmatrix}, \tag{2.60}$$

which is just the Lorentz transformation stated in (2.13).

Thus we see that $dx^\mu = (dx^0, dx^1, dx^2, dx^3) = (cdt, dx, dy, dz)$ form a 4-vector. Namely, these four components transform in a definite way under Lorentz transformation: observers in relative motion see different space and time components, or, as we say, space and time can transform into each other. Maxwell's electrodynamics equations have Lorentz symmetry because they are covariant under such transformations (instead of Galilean transformations). Thus, in order for Newtonian mechanics to be relativistic, that is, Lorentz symmetric, they must be generalized and reformulated. As we shall discuss in Chapter 10, we need to generalize the momentum from the familiar nonrelativistic expression of $\mathbf{p}_{\mathrm{NR}} = m\mathbf{v}$ to the relativistic momentum of $\mathbf{p} = \gamma m\mathbf{v}$ with γ again being the expression given in Eq. (2.14). In fact, observers in relative motion see different energy and momentum components. The three components of relativistic momentum, together with relativistic energy $E = \gamma mc^2$ form a 4-vector

$$p^\mu = (\gamma mc, \gamma mv_x, \gamma mv_y, \gamma mv_z) \tag{2.61}$$

$$= (p^0, p^1, p^2, p^3) \equiv \left(\frac{E}{c}, \mathbf{p}\right). \tag{2.62}$$

They transform into each other under Lorentz transformation in exactly the same manner as (cdt, dx, dy, dz) transform into each other. Just as the "length" of the 4-vector differential dx^μ is an Lorentz invariant ds^2 (cf. (2.48)), so is the length of the momentum 4-vector

$$\eta_{\mu\nu}p^{\mu}p^{\nu} = -\frac{E^2}{c^2} + p^2 = -m^2c^2. \tag{2.63}$$

Namely, the Lorentz invariant is $-m^2c^2$ with m being the rest mass, leading to the well-known relativistic energy–momentum relation

$$E = \sqrt{p^2c^2 + (mc^2)^2}. \tag{2.64}$$

More will be discussed of this relation in Chapter 10.

Given that Lorentz transformations may be viewed as "rotations" in the 4D spacetime, the physics equations written in terms of 4-vectors and 4-tensors, for example, as in (2.47) and (2.48), will automatically not change their form under Lorentz transformation—these equations will be, manifestly, relativistic. Tensors for such manifestly covariant formalism will be developed in Chapter 10 of Part III.

2.3.3 The spacetime diagram

Space and time coordinates are labels of physical processes taking place in the world. In this section, we discuss a particularly useful tool, the **space–time diagram**, to visualize this causal structure. To have the same length dimension for all coordinates, the temporal axis is represented, after using the conversion factor of light speed, by ct. The spacetime manifold is a representation of the relations among physical processes taking place in the world. In particular, with time being a coordinate, it reflects the causal structure of events taking place in the universe. The flat geometry of the spacetime in SR reflects the nature of physical relations that can be observed.

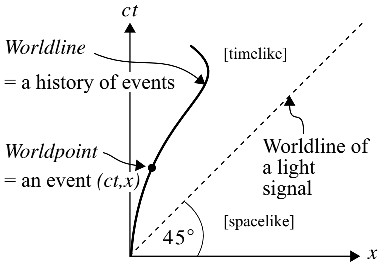

Fig. 2.6 Basic elements of a spacetime diagram, with two spatial coordinates suppressed.

Basic features and invariant regions

An event with coordinates (t, x, y, z) is represented by a **worldpoint** in the spacetime diagram. The history of events becomes a line of worldpoints, called a **worldline**. In Fig. 2.6 the three-dimensional space is represented only by an one-dimensional x-axis. In particular, a light signal passing through the origin $\Delta x = c\Delta t$ is represented by a straight worldline at a 45° angle with respect to the axes.

Since $\Delta s^2 = \Delta x^2 + \Delta y^2 + \Delta z^2 - c^2\Delta t^2$ is invariant, it is meaningful to divide the spacetime diagram into regions, as in Fig. 2.7, corresponding to

$\Delta s^2 < 0$	timelike
$\Delta s^2 = 0$	lightlike
$\Delta s^2 > 0$	spacelike

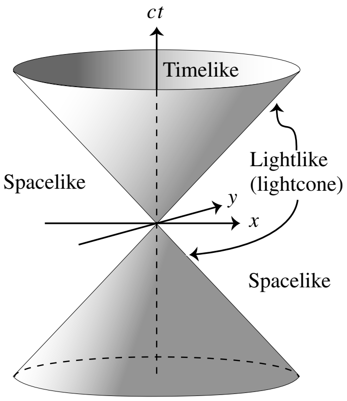

Fig. 2.7 Invariant regions in the spacetime diagram, with one spatial coordinates suppressed.

The coordinate intervals being $\Delta t = t_2 - t_1$, $\Delta x = x_2 - x_1$, etc., consider the separation of two events: one at the origin $(t_1, \mathbf{x}_1) = (0, \mathbf{0})$, the other at a point in one of the regions $(t_2, \mathbf{x}_2) = (t, \mathbf{x})$:

1. The lightlike region has all the events which are connected to the origin with a separation of $\Delta s^2 = 0$ hence for light signals. The 45° incline, called the **lightcone**, has a slope of unity which reflects the light speed being c.

2. The spacelike region has all the events which are connected to the origin with a separation of $\Delta s^2 > 0$. Namely, it will take a signal traveling at a speed greater than c in order to connect it to the origin. Thus, an event taking place at any point in this region cannot influence causally (in the sense of cause-and-effect) the event at origin, and vice versa. We can alternatively explain it by going to another frame S' resulting in a different spatial and time intervals $\Delta x' \neq \Delta x$ and $\Delta t' \neq \Delta t$. However, the spacetime interval is unchanged $\Delta s'^2 = \Delta s^2 > 0$. The form of (2.28) suggests that we can always find an S' frame such that this event would be viewed as taking place at the same time $\Delta t' = 0$ as the event at the origin but at different locations $\Delta x' \neq 0$. This makes it clear that such a worldpoint (an event) **cannot be causally connected** to the origin.

3. The timelike region has all the events, which are connected to the origin with a separation of $\Delta s^2 < 0$. One can always find a frame S' such that such an event takes place at the same locations $x' = 0$ but at different time $t' \neq 0$. This makes it clear that it **can be causally connected** with the origin. In fact, all the worldlines passing through the origin will be confined in this region, inside the lightcone. (*Remark:* The worldline of an inertial observer must be a straight line because of constant velocity inside the lightcone. These straight lines are just the time axes of the coordinate systems in which the inertial observer is at rest.)

In Fig. 2.7 we have displayed the lightcone structure with respect to the origin of the spacetime coordinates ($t = 0$, $\mathbf{x} = 0$). It should be emphasized that each point in a spacetime diagram has a lightcone. The timelike regions with respect to worldpoints $P_1, P_2, \ldots$ as represented by the lightcones at $P_1, P_2, \ldots$ are shown in Fig. 2.8.

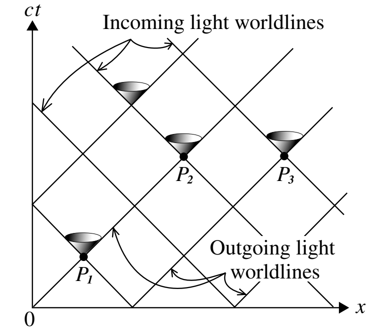

Fig. 2.8 Lightcones with respect to different worldpoints, $P_1, P_2, \ldots$, etc.

Lorentz transformation in the spacetime diagram

The nontrivial parts of the Lorentz transformation of intervals (taken with respect to the origin) being

$$\Delta x' = \gamma(\Delta x - \beta c \Delta t), \quad c\Delta t' = \gamma(c\Delta t - \beta \Delta x) \tag{2.65}$$

the transformed axes are:

1. The x'-axis corresponds to the $ct' = 0$ line. Hence it is a straight line in the (x, ct) plane with a slope of β.
2. The ct'-axis corresponds to the $x' = 0$ line. Hence it is a straight line with a slope of β^{-1}.

Depending on whether β is positive or negative, the new axes "close-in" or "open-up" from the original perpendicular axes. Thus we have the **opposite-angle rule**: the two axes make opposite-signed $\pm\theta$ rotations, Fig. 2.9. (The x-axis rotates by $+\theta$ to the x'-axis; the ct-axis by $-\theta$ to the ct'-axis.) The physical basis for this rule is the requirement to maintain the same slope (=1, that is, **equal angles** with respect to the two axes) for the lightcone in every inertial frame.

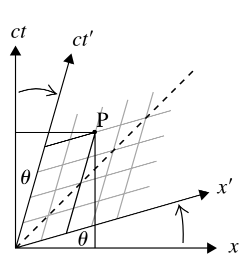

Fig. 2.9 Lorentz rotation in the spacetime diagram. The space and time axes rotate for the same amount, in opposite directions, so that the lightcone (the dashed line) remains unchanged. The shaded grid represents lines of fixed x' and t'. What's displayed is for $\beta > 0$ with axes "closing-in."

Another important feature of the diagrammatic representation of the Lorentz transformation is that the new axes will have a scale **different** from the original one. Namely, the unit-length along the axes of the two systems are different. Let us illustrate this by an example.

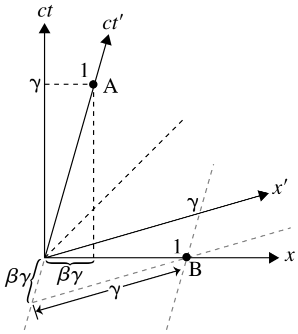

Fig. 2.10 Scale change in a Lorentz rotation. For example, a unit length on the ct' axis has a **longer** projection γ onto the ct axis. Namely, the event A ($ct' = 1, x' = 0$) is observed by O to have the coordinates ($ct = \gamma, x = \gamma\beta$). Similarly, the event B with ($ct = 0, x = 1$) has the coordinates ($ct' = \gamma\beta, x' = \gamma$). The two sets of dotted lines passing through worldpoints A and B are parallel lines to the axes of (ct, x) and to (ct', x'), respectively.

Consider the separation (from origin O) of an event A on the ct' axis, which has O' system coordinates ($ct' = 1, x' = 0$), see Fig. 2.10. What O system coordinates (ct, x) does the worldpoint have?

$$x' = \gamma(x - \beta ct) = 0 \Rightarrow x = \beta ct,$$

$$ct' = \gamma(ct - \beta x) = ct\gamma(1 - \beta^2) = \frac{ct}{\gamma} = 1.$$

Hence this event has ($ct = \gamma, x = \gamma\beta$) coordinates in the O system. Evidently, as $\gamma > 1$, a unit vector along the ct' direction has the "projection" on the ct-axis that is longer than unit length. This is possible only if there is a scale change when transforming from one reference system to another.

Consider another separation of an event B on the x-axis, which has O coordinates ($ct = 0, x = 1$). It is straightforward to check that it has O' system coordinates ($ct' = -\gamma\beta, x' = \gamma$), again showing a difference in scales of the two systems.

2.3.4 Time-dilation and length contraction

What is the physics behind the above-discussed scale changes? The answer, to be presented in the following subsections, is **time-dilation** and **length-contraction**:

> A moving clock **appears** to run slow,
> a moving object **appears** to contract.

These physical features underscore the profound change in our conception about space and time brought on by relativity. We must give up our belief that measurements of space and time give the same results for all observers. Special relativity makes the strange claim that observers in relative motion will have different perceptions of distance and time. This means that two identical watches worn by two observers in relative motion will tick at different rates and will not agree on the amount of time that has elapsed between two given events. It is not that these two watches are defective. Rather, it is a fundamental statement about the nature of time.

While the algebra involved in deriving these results (from Lorentz transformation) is simple, to obtain the correct result, one has to be very clear in **what measurements** precisely are being compared in two different frames.

Time-dilation

A clock, ticking away in its own rest frame O' (also called the comoving frame), is represented by a series of worldpoints (the ticks) equal-spaced on a vertical worldline ($\Delta x' = 0$) in the ct'-x' spacetime diagram. These same worldpoints when viewed in another inertial frame O in which the O' system moves with $+v$ along the x-axis will appear as lying on an inclined worldline, Fig. 2.11.

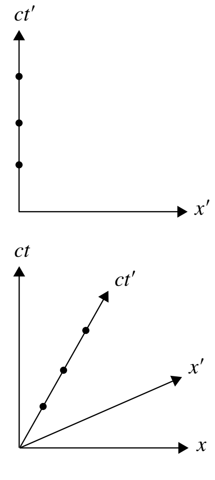

Fig. 2.11 Worldline of a clock, ticking at equal intervals: viewed in the rest frame of the clock, the O' system, and viewed in the moving frame, the O-coordinate system.

From (2.65), as well as our previous discussion of the scale change under Lorentz rotation (also see Fig. 2.13), it is clear that these intervals will be measured to have interval

$$\Delta t = \gamma \Delta t' \qquad \gamma > 1. \tag{2.66}$$

Thus, we say that a moving clock (i.e. moving with respect to the O system) appears (Δt) to run slow. (NB: Keep in mind that the comoving frame has $\Delta x' = 0$ while the moving frame $\Delta x \neq 0$.)

Physically there is an easy way to understand this phenomenon of time dilation. Stripping away all extraneous mechanisms, consider the most basic of clocks[8]: a light-pulse clock. It ticks away the time by having a light-pulse bouncing back and forth between a fixed distance d (Fig. 2.12).

For a comoving observer, one has

$$\Delta t' = \frac{d}{c}. \tag{2.67}$$

To an observer with respect to whom the clock is moving with speed v (say, perpendicular to the light-pulse path) the light-pulse will traverse a diagonal distance D at a different time interval Δt

$$\Delta t = \frac{D}{c} = \frac{\sqrt{d^2 + v^2 \Delta t^2}}{c}. \tag{2.68}$$

Collecting Δt terms, we have

$$\Delta t = \frac{d/c}{\sqrt{1 - v^2/c^2}} = \gamma \Delta t' \tag{2.69}$$

showing the time-dilation result of (2.66).

Time-dilation seems counter-intuitive. This is so because our intuition has been built up from familiar experience with phenomena having velocity much less than c. Actually, it is easier to understand such physical results at an extreme speed regime of $v \lesssim c$. Let us look at this phenomenon in a situation having $v = c$. In this case, time is infinitely dilated. Imagine a rocket-ship traveling at $v = c$ passing a clock tower. The rocket pilot (the observer in O frame) will see the clock (at rest in the O' frame) as infinitely dilated (i.e. stopped) at the instant when the ship passes the tower. This just means that the light image of the clock cannot catch up with the rocket-ship.

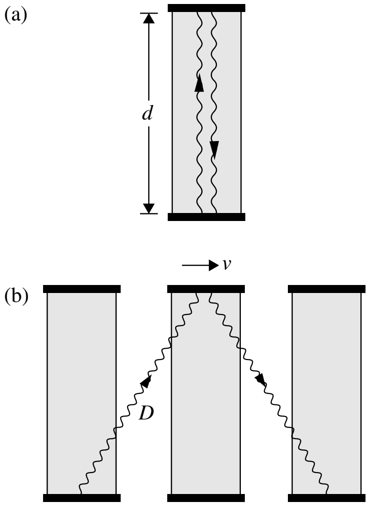

Fig. 2.12 Light-pulse clock at rest (a) and in motion (b).

Length contraction

The length of a moving object Δx, compared to the $\Delta x'$ as measured in its own rest frame O', appears to be shortened—often called the **FitzGerald–Lorentz contraction** in the literature.

To obtain a length $\Delta x = x_1 - x_2$ in the O system, we need to measure two events (t_1, x_1) and (t_2, x_2) simultaneously $\Delta t = t_1 - t_2 = 0$. (If you want to measure the length of a moving car, you certainly would not want to measure its front and back locations at different times!) The same two events, when viewed in the rest frame of the object$\Delta x' = x_1' - x_2'$, will be measured according to

[8] Different clocks—mechanical clocks, biological clocks, atomic clocks, or particle decays via strong or weak interactions—simply represent different physical phenomena that can be used to mark time. A "basic clock" rests on some physical phenomenon that has a direct connection to the underlying physics laws.

(2.65) to have a different separation (cf. Fig. 2.13)

$$\Delta x' = \gamma \Delta x > \Delta x. \tag{2.70}$$

(NB: While we have simultaneous measurements in the O-system, $\Delta t = 0$, these two events would be viewed as taking place at different times in the O'-system, $\Delta t' = \gamma(\Delta t - (v/c^2)\Delta x) \neq 0$. Of course, in the rest frame of the object, there is no need to perform the measurements simultaneously—to measure the front and back ends of a parked car it is perfectly all right to make one measurement, take a lunch break, and come back to measure the other end.)

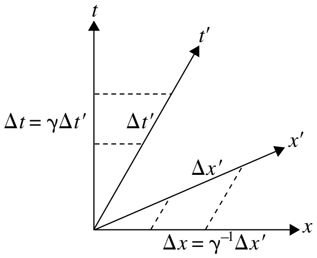
t
t'
$\Delta t = \gamma \Delta t'$
$\Delta t'$
x'
$\Delta x'$
x
$\Delta x = \gamma^{-1} \Delta x'$

Fig. 2.13 Scale change of the Lorentz rotation reflecting the physics phenomena of time-dilation and length contraction. The clock and object are moving with respect to the O-system, but are at rest with respect to the O'-system.

Length contraction, observed from the O-system, is only in the direction of relative motion of the frames. Thus the volume contraction is same as length contraction: $V = \gamma^{-1} V'$, and not $\gamma^{-3} V'$.

Physical interpretation of the terms in the Lorentz transformation

We have used (2.65) to deduce the phenomena of time-dilation and length contraction. Now we will revert the reasoning to find the physical interpretation of the terms in (2.65).

In the space transformation

$$\Delta x' = \gamma(\Delta x - v\Delta t), \tag{2.71}$$

the $(\Delta x - v\Delta t)$ factor is the same as in the familiar Galilean transformation. The overall factor γ simply reflects the physics of length contraction. In particular for the length measurement (thus $\Delta t = 0$), the relation $\Delta x' = \gamma \Delta x$ says that the O-system length Δx is shorter than the O'-system length $\Delta x'$ by a factor of γ.

The time transformation

$$\Delta t = \gamma\left(\Delta t' + \frac{v}{c^2}\Delta x'\right) \tag{2.72}$$

has two terms. The first term represents the time-dilation effect: $\Delta x' = 0$ in the O'-system with $\Delta t'$ being the proper time interval; the second term is the amount of nonsynchronization that has developed in the O-system, between two clocks, located at different positions $\Delta x' \neq 0$ in the O' frame. The two clocks are synchronized $(\Delta t' = 0)$ in the O' frame. However, for the observer in the O system, there will be a lack of simultaneity, according to (2.72), equal to $\Delta t = \gamma\beta\Delta x'/c$ (Fig. 2.14). This indeed agrees with the result obtained in Eq. (2.27) when we first discussed relativity of simultaneity (in Section 2.2.2 and Fig. 2.4) with the rail-car observer seeing the distance between the two bulbs (length of the car) as $L_{\rm c} = \Delta x'$ while the same length would appear to the platform observer as being $L_{\rm p} = \gamma^{-1}\Delta x'$ due to length contraction. NB the distance separation between the two events of light emissions, as seen by the platform observer, would be $\Delta x = \gamma \Delta x'$ (see Problem 2.15).

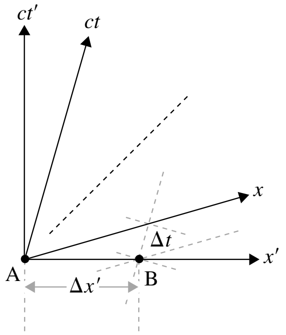

Fig. 2.14 Relativity of simultaneity (cf. Section 2.2.2). Two events A (located at the origin) and B (located at $\Delta x'$) are simultaneous $\Delta t' = 0$ as viewed by the observer O'. With respect to a moving observer O, they are no longer simultaneous, their coordinate time difference is Δt. From the scale change as displayed in Fig. 2.10 we have $c\Delta t = \gamma\beta\Delta x'$.

Twin paradox We also refer the reader to Section A.1 for a discussion of the "twin paradox," which is a particularly instructive example that illuminates several basic concepts in relativity.

Review questions

1. Two inertial frames are moving with respect to each other with velocity $\mathbf{v} = v\,\hat{\mathbf{x}}$. Write out the Lorentz transformation of coordinates $(t, \mathbf{x})$? Show that in the low velocity limit it reduces to Galilean transformation.

2. Under a Lorentz transformation, the electric and magnetic fields transform into each other. Give a simple physical explanation of a situation when a static electric field between two charges gives rise to magnetic field when viewed by a moving observer.

3. Under Lorentz transformation, not only the spatial interval (Δx for $\Delta t \neq 0$), but also the time interval (Δt for $\Delta x \neq 0$), are relative. What combination of spatial and time intervals remain to be absolute? What does one mean by relative and absolute in this context? Give a simple physical explanation why we expect this combination to be absolute. How can this statement be interpreted geometrically as showing that the Lorentz transformation is a generalized rotation in a 4D pseudo-Euclidean flat manifold?

4. Lorentz transformation was first discovered as some mathematical property of Maxwell's equations. What did Einstein do to provide it with a physical basis?

5. Give the definition of the metric tensor in terms of (a) the basis vectors, and (b) infinitesimal length separation. When the metric is displayed as a square matrix, what is the respective interpretation of its diagonal and off-diagonal elements? What is the metric for a Cartesian coordinate system?

6. What is the condition on the rotation matrix $[\mathbf{R}]$ that reflects, for a general coordinate system, the length preserving nature of a rotation transformation in a flat space (with a metric $[\mathbf{g}]$)? Check that this equation reduces to the familiar orthogonality condition for the case of Cartesian coordinates.

7. From the condition expressing the invariance of the metric, derive the explicit form of the Lorentz transformation for a boost $\mathbf{v} = +v\,\hat{\mathbf{x}}$.

8. In the spacetime diagram, display the timelike, spacelike, and lightlike regions. Also, draw in the worldline for some inertial observer.

9. What does one mean by the saying that "simultaneity is relative"? Illustrate this in a spacetime diagram.

10. The coordinate frame O' is moving at a constant velocity v in the $+x$ direction with respect to the O coordinate frame. Display the transformed axes (x', ct') in a two-dimensional spacetime diagram with axes (x, ct). (You are not asked to solve the Lorentz transformation equations, but rather to justify the directions of the new axes.)

11. **Length contraction** means that the measured length interval of $\Delta x = x_1 - x_2$ is less than the corresponding rest-frame length $\Delta x' = x'_1 - x'_2$. What is the condition on the time coordinates of these two events: (t_1, t_2) and (t'_1, t'_2)? Use this condition and the Lorentz transformation to derive the result of length contraction.

12. **Time-dilation** means that the measured time interval of $\Delta t = t_1 - t_2$ is longer than the corresponding rest-frame interval $\Delta t' = t'_1 - t'_2$. What is the condition on the spatial coordinates of these two events: (x_1, x_2) and (x'_1, x'_2)? Use this condition and the Lorentz transformation to derive the result of time-dilation.

13. Draw the lines of simultaneity $\Delta t = 0$ (i.e. lines on which all worldpoints have the same time coordinate) and the lines of spatial congruity $\Delta x = 0$ in the two-dimensional (x, ct) spacetime diagram, as well as the $\Delta t' = 0$ and $\Delta x' = 0$ lines in the same diagram.

Problems

(2.1) **Newtonian relativity** Consider a few definite examples of Newtonian mechanics that are unchanged by Galilean transformation of (2.10) and (2.11):

(a) Show that force as given by the product of acceleration and mass $\mathbf{F} = m\mathbf{a}$, as well as a force law, for example, Newton's law of gravity $\mathbf{F} = G_{\mathrm{N}}(m_1 m_2 / r^2)\hat{\mathbf{r}}$, remain the same in every inertial frame;

(b) **Energy and momentum conservation** laws hold in different inertial frames. Consider the one-dimensional collision having initial (**mass,velocity**) of two particles, in some appropriate units, being $(3, 8)$ and $(5, -4)$ and final configuration of $(3, -7)$ and $(5, 5)$. Check that energy and momentum are conserved when viewed by observers in both reference frames having a relative speed of 2. Also work out the same conservation law for the general case of $(m_1, v_1) + (m_2, v_2) \rightarrow (m_1, v_1') + (m_2, v_2')$ with u being the relative velocity of the two observers.

(2.2) **Inverse Lorentz transformation** The Lorentz transformation of (2.13) may be written in the matrix form as

$$\begin{pmatrix} ct' \\ x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} \gamma & -\beta\gamma & 0 & 0 \\ -\beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix}. \qquad (2.73)$$

Find the inverse transformation, that is, find the coordinates (ct, x, y, z) in terms of (ct', x', y', z'), by the physical expectation of the inverse being given by changing the sign of the relative velocity v. Show that the transformation found in this way is indeed inverse to the matrix in (2.73) by explicit matrix multiplication.

(2.3) **Lorentz transformation of derivative operators** The Lorentz transformation as shown in Eq. (2.73) can be written in component notion as (the Greek indices $\mu = 0, 1, 2, 3$ with $x^0 = ct$).

$$x'^{\mu} = \sum_{\nu} [\mathbf{L}]^{\mu}_{\nu} x^{\nu}. \qquad (2.74)$$

Here we seek the transformation $[\bar{\mathbf{L}}]$ for the coordinate derivatives

$$\partial'_{\mu} = \sum_{\nu} [\bar{\mathbf{L}}]_{\mu}^{\ \nu} \partial_{\nu}, \qquad (2.75)$$

where

$$\partial_{\mu} = \left(\frac{\partial}{c\partial t} \quad \frac{\partial}{\partial x} \quad \frac{\partial}{\partial y} \quad \frac{\partial}{\partial z} \right).$$

Show that the coordinate derivative operators $((\partial/\partial t),\ (\partial/\partial x_i))$ transform oppositely $(v \rightarrow -v)$ compared to the coordinate (t, x_i) transformation (2.13):

$$\begin{aligned} \frac{\partial}{\partial t'} &= \gamma\left(\frac{\partial}{\partial t} + v\frac{\partial}{\partial x} \right), \\ \frac{\partial}{\partial x'} &= \gamma\left(\frac{\partial}{\partial x} + \frac{v}{c^2}\frac{\partial}{\partial t} \right), \qquad (2.76) \\ \frac{\partial}{\partial y'} &= \frac{\partial}{\partial y}, \quad \frac{\partial}{\partial z'} = \frac{\partial}{\partial z}. \end{aligned}$$

Namely, in terms of the transformation matrices, $[\bar{\mathbf{L}}] = [\mathbf{L}]^{-1}$.

(a) Obtain this result by the standard chain rule of differentiation.

(b) Obtain this result by the observation that $\partial x^{\nu}/\partial x^{\mu} = \delta^{\mu\nu}$ is an invariant.

(2.4) **Lorentz covariance of Maxwell's equations** Show that Maxwell's equation and Lorentz force law are covariant under the Lorentz transformation as given in (2.76), (2.18), and (2.19). Suggestion: Work with the homogeneous and nonhomogeneous parts of the Maxwell's equations separately. For example, show that

$$\nabla' \cdot \mathbf{B}' = 0, \quad \nabla' \times \mathbf{E}' + \frac{1}{c}\frac{\partial \mathbf{B}'}{\partial t'} = 0, \qquad (2.77)$$

follows, by Lorentz transformation, from

$$\nabla \cdot \mathbf{B} = 0, \quad \nabla \times \mathbf{E} + \frac{1}{c}\frac{\partial \mathbf{B}}{\partial t} = 0. \qquad (2.78)$$

(2.5) **From Coulomb's to Ampere's law** Just as we can derive $F_y = ma_y$ and $F_z = ma_z$ from $F_x = ma_x$ by rotation transformations, show that, by Lorentz transformations, one can derive (a) Ampere's law (with conduction and displacement currents) from Coulomb's/Gauss's law, and vice versa; (b) Magnetic Gauss's law (absence of magnetic monopole) follows from Faraday's law of induction, and vice versa.

(2.6) **The Lorentz invariant space–time interval** Use the explicit form of the Lorentz transformation given in (2.13) to show that $s^2 = x^2 + y^2 + z^2 - c^2t^2$ is an invariant.

(2.7) **Rotation matrix is orthogonal** Explicitly demonstrate that the rotation matrix in Eq. (2.54) satisfies the relation

$$[\mathbf{R}^{-1}(\theta)] = [\mathbf{R}(-\theta)] = [\mathbf{R}^{\top}(\theta)]$$

hence the orthogonality condition $\mathbf{R}\mathbf{R}^{\top} = \mathbf{1}$.

(2.8) **Group property of Lorentz transformations** Use simple trigonometry to show that the rotation and boost operators given in (2.54) and (2.57) satisfy the group property:

$$\begin{aligned} [\mathbf{R}(\theta_1)][\mathbf{R}(\theta_2)] &= [\mathbf{R}(\theta_1 + \theta_2)], \\ [\mathbf{L}(\psi_1)][\mathbf{L}(\psi_2)] &= [\mathbf{L}(\psi_1 + \psi_2)]. \end{aligned} \qquad (2.79)$$

(2.9) **Transformation multiplication leads to velocity addition rule** Provide a proof of the velocity addition rule (2.24) by way of (2.79) in Problem 2.8.

(2.10) **Spacetime diagram depicting relativity of simultaneity** Draw two spacetime diagrams (a) one showing the worldlines of an observer seeing two bolts of lightning taking place simultaneously, (b) another diagram depicting the viewpoint of a moving observer seeing these two lightning bolts as taking place at different times. (You are asked to draw three worldlines in each of these two diagrams: one for the observer, two for the lightning.)

(2.11) **Length contraction and light-pulse clock** In Section 2.3.4 we have used a light-pulse clock to demonstrate the phenomenon of time-dilation. This same clock can be used to demonstrate length contraction: the length of the clock l can be measured through the time interval Δt that takes a pulse to make the trip across the length of the clock and back: $2l = c\Delta t$. Deduce the length contraction formula (2.70) in this setup. Suggestion: In the case of time-dilation, we had the clock moving in the direction perpendicular to the rest-frame light path. Here you want it to be parallel. Also, you will need to use the time-dilation formula (2.66) to deduce the final length contraction result.

(2.12) **Pion decay-length in the laboratory** In high energy proton–proton collisions, copious number of (subatomic particle) pions are produced. Even though pions have a half life time of $\tau_0 = 1.77 \times 10^{-8}$ s (they decay into a final state of a muon and a neutrino via weak interaction), they can be collimated to form pion beams for other high energy physics experiments. This is possible because the pions produced from pp collisions have high kinetic energy, hence high velocity. For example, if the beam of pions has a velocity $0.99c$, it can retain half of its intensity even after traveling a distance close to 38 m. This may be surprising because a naive calculation of $\tau_0 c = (1.77 \times 10^{-8}) \times (3 \times 10^8) = 5.3\,\text{m}$ would lead one to expect a much shorter decay length. Explain why the naive calculation is incorrect. Perform the correct calculations by using (a) time-dilation, and (b) length contraction. Explain clearly which reference frame one uses to get these results.

(2.13) **Two spaceships passing one another** Two spaceships traveling in opposite directions pass one another at a relative speed of 1.25×10^8 m/s. The clock on one spaceship records a time duration of 9.1×10^{-8} s for it to pass from the front end to the tail end of the other ship. What is the length of the second ship as measured in its own rest frame?

(2.14) **Invariant spacetime interval and relativity of simultaneity** Two events are spatially apart $\Delta x \neq 0$ but simultaneous $\Delta t = 0$ in one frame. When viewed in another inertial frame, they are no longer seen as taking place at the same time $\Delta t' \neq 0$. Find this time separation $\Delta t'$ in terms of Δx and $\Delta x'$ in two ways: (a) using the invariant spacetime interval, and (b) using the Lorentz transformation.

(2.15) **More simultaneity calculations** Work out the spacetime coordinates (x, t)'s of the two light emission events located at opposite ends of a moving rail-car—as seen by the observer on the car and by a platform observer as described in Section 2.2.2, cf. Fig. 2.4.

(a) Let the O' coordinates be the rail-car observer system, and O the platform observer system. Given $\Delta t' = 0$, use Lorentz transformation and its inverse to find the relations between Δt and Δx, between Δt and $\Delta x'$, and between Δx and $\Delta x'$.

(b) One of the relations obtained in (a) should be $\Delta x = \gamma \Delta x'$. Is this compatible with the derivation of length contraction as done in Section 2.4? Explain.

(c) An observer can locate the time the light emission took place by calculating the time it took the light signal to reach the observer. If the interval is t_1 then it must be emitted at time $-t_1$. (Namely, we define the arrival time as being $t = 0$.) By the same token, the emission must have taken place at a location $x_1 = -ct_1$. In this way, verify the relation $\Delta x = \gamma \Delta x'$ discussed earlier.

(d) Draw two sets of spacetime diagrams. (i) In one set show two light bulbs on the x'-axis, equidistant from origin, emit light pulses, which are received by the standstill-observer at the origin. Depict the same events according to a moving observer: light emitted from two points equidistant on the x-axis with the light ray not meeting at the same point on the t' axis. (ii) In another set of spacetime diagrams, show the two light emitting events as seen by one observer to be simultaneous and yet according to another moving observer, to be not so.

3 The principle of equivalence

- After a review of the Newtonian theory of gravitation in terms of its potential function, we take the first step in general relativity (GR) study with the introduction of the equivalence principle (EP).
- The *Weak* EP (the equality of the gravitational and inertial masses) is extended by Einstein to the *Strong* EP. This implies the existence of *local inertial frames* at every spacetime point. In a sufficiently small region, the "local inertial observer" will not sense any gravity effect.
- The equivalence of acceleration and gravity means that GR (physics laws valid in all coordinate systems, including accelerating frames) must necessarily be a theory of gravitation.
- The strong EP is used to deduce the results of gravitational bending of a light ray, gravitational redshift, and time dilation.
- Einstein was motivated by EP physics to propose a curved spacetime description of the gravitational field.

Soon after completing his formulation of special relativity (SR) in 1905, Einstein started working on a relativistic theory of gravitation. In this chapter, we cover the period of 1907–1911, when Einstein relied heavily on the equivalence principle (EP) to extract some general relativity (GR) results. Not until the end of 1915 did he work out fully the ideas of GR. By studying the consequences of EP, he concluded that proper language for GR is Riemannian geometry. The mathematics of curved space will be introduced in Chapter 4 and the geometric representation of gravitational field in Chapter 5.

3.1 Newtonian gravitation potential—a review

Newton formulated his theory of gravitation through the concept of action-at-a-distance force

$$\mathbf{F}(\mathbf{r}) = -G_\mathrm{N}\frac{mM}{r^2}\hat{\mathbf{r}}, \tag{3.1}$$

where G_N is **Newton's constant**, the point mass M is located at the origin of the coordinate system, and m is at position $\mathbf{r}$.

Just as the case of electrostatics $\mathbf{F}(\mathbf{r}) = q'\mathbf{E}(\mathbf{r})$, we can cast this in the form

$$\mathbf{F}(\mathbf{r}) = m\mathbf{g}(\mathbf{r}). \tag{3.2}$$

This defines the gravitational field $\mathbf{g}(\mathbf{r})$ as the gravitational force per unit mass. Newton's law, in terms of this gravitational field for a point mass M, is

$$\mathbf{g}(\mathbf{r}) = -G_\mathrm{N}\frac{M}{r^2}\hat{\mathbf{r}}. \tag{3.3}$$

Just as Coulomb's law can be equivalently stated as Gauss's law for the electric field, this field Eq. (3.3) can be expressed, for an arbitrary mass distribution, as Gauss's law for the gravitational field:

$$\oint_{S} \mathbf{g} \cdot d\mathbf{A} = -4\pi G_{\mathrm{N}} M. \tag{3.4}$$

The area integral on the left-hand side (LHS) is the gravitational field flux through any closed-surface S, and M on the right-hand side (RHS) is the total mass enclosed inside S. This integral representation of Gauss's law (3.4) can be converted into a differential equation: we will first turn both sides into volume integrals by using the divergence theorem on the LHS (the area integral into the volume integral of the divergence of the field) and by expressing the mass on the RHS in terms of the mass density function ρ

$$\int \nabla \cdot \mathbf{g}\, dV = -4\pi G_{\mathrm{N}} \int \rho\, dV.$$

Since this relation holds for any volume, the integrands on both sides must also equal:

$$\nabla \cdot \mathbf{g} = -4\pi G_{\mathrm{N}} \rho. \tag{3.5}$$

This is Newton's field equation in differential form. Gravitational potential[1] $\Phi(x)$ being defined through the gravitational field $\mathbf{g}(x) \equiv -\nabla\Phi(x)$, the field Eq. (3.5) becomes

$$\nabla^2 \Phi = 4\pi G_{\mathrm{N}} \rho. \tag{3.6}$$

To obtain the gravitational equation of motion, we insert (3.2) into Newton's second law $\mathbf{F} = m\ddot{\mathbf{r}}$,

$$\ddot{\mathbf{r}} = \mathbf{g}, \tag{3.7}$$

which has the outstanding feature of being totally independent of any properties (mass, charge, etc.) of the test particle. Expressed in terms of the gravitational potential, it can now be written as

$$\ddot{\mathbf{r}} = -\nabla\Phi. \tag{3.8}$$

We note that the Newtonian field theory of gravitation, as embodied in (3.6) and (3.8), is not compatible with SR as space and time coordinates are not treated on equal footings. In fact Newtonian theory is a **static** field theory. Stated in another way, these equations are comparable to Coulomb's law in electromagnetism. They are not complete, as the effects of motion (i.e. magnetism) are not included. This "failure" just reflects the underlying physics that only admits an (instantaneous) action-at-a-distance description, which implies an infinite speed of signal transmission, incompatible with the principle of relativity.

3.2 EP introduced

In this section, several properties of gravitation will be presented. They all follow from the empirical principle called by Einstein the **principle of the equivalence of gravitation and inertia**. The final formulation of Einstein's theory of gravitation, the general theory of relativity, automatically and precisely contains this EP. Historically, it is the starting point of a series of discoveries that ultimately led Einstein to the geometric theory of gravity, in which the gravitation field is the warped spacetime.

[1] We have the familiar example of potential for a spherically symmetric source with total mass M given by $\Phi = -G_{\mathrm{N}}M/r$.

3.2.1 Inertial mass vs. gravitational mass

One of the distinctive features of gravitation field is that its equation of motion (3.8) is totally independent of the test particle's properties (mass, charge, etc.). This comes about because of the cancellation of the mass factors in $m\mathbf{g}$ and $m\ddot{\mathbf{r}}$. Actually, these two masses correspond to very different concepts:

- The **inertial mass**
$$\mathbf{F} = m_{\mathrm{I}}\ddot{\mathbf{r}} \tag{3.9}$$
enters into the description of the response of a particle to **all** forces.
- The **gravitational mass**
$$\mathbf{F} = m_{\mathrm{G}}\mathbf{g} \tag{3.10}$$
reflects the response[2] of a particle to a **particular** force: gravity. The gravitational mass m_{G} may be viewed as the "gravitational charge" placed in a given gravitational field $\mathbf{g}$.

Now consider two objects A and B composed of different material, one of copper and the other of wood. When they are let go in a given gravitational field $\mathbf{g}$, e.g. "being dropped from the Leaning Tower of Pisa" (see Box 3.1), they will, according to (3.9) and (3.10), obey the equations of motion:

$$(\ddot{\mathbf{r}})_{\mathrm{A}} = \left(\frac{m_{\mathrm{G}}}{m_{\mathrm{I}}}\right)_{\mathrm{A}} \mathbf{g}, \quad (\ddot{\mathbf{r}})_{\mathrm{B}} = \left(\frac{m_{\mathrm{G}}}{m_{\mathrm{I}}}\right)_{\mathrm{B}} \mathbf{g}. \tag{3.11}$$

Part of Galileo's great legacy to us is the **experimental observation** that all bodies fall with the same acceleration, $(\ddot{\mathbf{r}})_{\mathrm{A}} = (\ddot{\mathbf{r}})_{\mathrm{B}}$, which leads to the equality,

$$\left(\frac{m_{\mathrm{G}}}{m_{\mathrm{I}}}\right)_{\mathrm{A}} = \left(\frac{m_{\mathrm{G}}}{m_{\mathrm{I}}}\right)_{\mathrm{B}}. \tag{3.12}$$

The mass ratio, having been found to be universal for all substances as in (3.12), can then be set, by appropriate choice of units, equal to unity. This way we can simply say

$$m_{\mathrm{I}} = m_{\mathrm{G}}. \tag{3.13}$$

Even at the fundamental particle physics level, matter is made up of protons, neutrons, and electrons (all having different interactions) bound together with different binding energies, it is difficult to find an a priori reason to expect such a relation (3.12). As we shall see, this is the empirical foundation underlying the geometric formulation of relativistic theory of gravity that is GR.

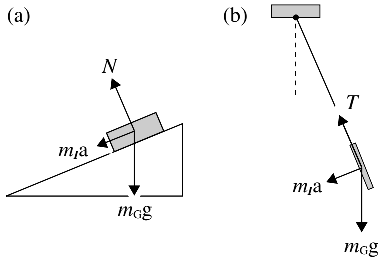

Fig. 3.1 Both the gravitational mass and inertia mass enter into the phenomena: (a) sliding object on an inclined plane, and (b) oscillations of a pendulum.

Box 3.1 A brief history of the EP: from **Galileo** and **Newton** to **Eötvös**

There is no historical record of Galileo having dropped anything from the Leaning Tower of Pisa. Nevertheless, to refute Aristotle's contention that heavier objects would fall faster than light ones, he did report performing experiments of sliding different objects on an inclined plane, Fig. 3.1(a). (The slower fall allows for more reliable measurements.) More importantly, Galileo provided a theoretical argument, "a thought experiment," in the first chapter of his "**Discourse and Mathematical Demonstration of Two New Sciences**," in support of the idea that all substances should fall with the same acceleration. Consider any falling object, without this universality of free

[2] One should in principle distinguish between two separate gravitational charges: a "passive" gravitational mass as in (3.10) is the response to the gravitational field, and an "active" gravitational mass is the source of gravitational field $\mathbf{g} = -G_{\mathrm{N}}m_{\mathrm{G}}\mathbf{r}/r^3$. These two masses can be equated by way of Newton's third law.

fall, the tendency of different components of the object to fall differently would give rise to internal stress and could cause certain objects to undergo spontaneous disintegration. The nonobservation of this phenomenon could then be taken as evidence for equal accelerations.

Newton went further by translating this universality of free fall into the universal proportionality of the inertial and gravitational masses (3.12) and built the equality $m_{\rm I} = m_{\rm G}$ right into the foundation of mechanics. Notably, he discussed this equality in the very first paragraph of his great work "**Principia**." Furthermore, he improved upon empirical check of (3.12) of Galileo by experimenting with a pendulum, Fig. 3.1(b), cf. Problem 3.1,

$$\delta_{\rm AB} \equiv \left| \frac{(m_{\rm I}/m_{\rm G})_{\rm A} - (m_{\rm I}/m_{\rm G})_{\rm B}}{(m_{\rm I}/m_{\rm G})_{\rm A} + (m_{\rm I}/m_{\rm G})_{\rm B}} \right| \leq 10^{-3}. \qquad (3.14)$$

The Eötvös experiment and modern limits. At the end of the nineteenth century, the Hungarian baron Roland von Eötvös pointed out that any possible nonuniversality of this mass ratio (3.12) would show up as a horizontal twist τ in a torsion balance, Fig. 3.2(b). Two weights composed of different substances A and B are hung at the opposite ends of a rod, which is in turn hung from the ceiling by a fiber at a midpoint, respective distances $l_{\rm A}$ and $l_{\rm B}$ from the two ends. Because of Earth's rotation, we are in a noninertial frame of reference. In order to apply Newton's laws, we must include the fictitious force, as represented by the centrifugal acceleration $\mathbf{g}'$, Fig. 3.2(a). In the vertical direction we have the gravitational acceleration g, and the (tiny and, for our simplified calculation, negligible) vertical component $g'_{\rm v}$. In the horizontal direction the only nonzero torque is due to the horizontal component $g'_{\rm h}$. The equilibrium conditions of a vanishing total torque are:

$$\text{vertical balance:} \qquad [l_{\rm A}(m_{\rm G})_{\rm A} - l_{\rm B}(m_{\rm G})_{\rm B}]g = 0, \qquad (3.15)$$

$$\text{horizontal balance:} \qquad [l_{\rm A}(m_{\rm I})_{\rm A} - l_{\rm B}(m_{\rm I})_{\rm B}]g'_{\rm h} = \tau. \qquad (3.16)$$

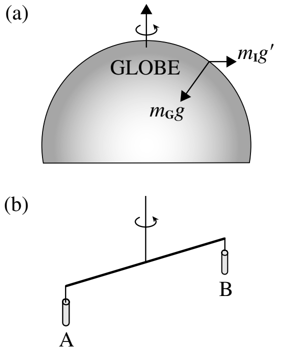

Fig. 3.2 Eötvö s experiment to detect any difference between the ratio of gravitational to inertial masses of substance A vs. B. The centrifugal acceleration can be decomposed into the vertical and horizontal components, $\mathbf{g}' = g'_{\rm v} + g'_{\rm h}$.

The equality of $l_{\rm A}(m_{\rm G})_{\rm A} = l_{\rm B}(m_{\rm G})_{\rm B}$ from the equilibrium condition of (3.15) means that the twist in (3.16) is related to the sought-after nonuniversality:

$$\tau = \left[\left(\frac{m_{\rm I}}{m_{\rm G}}\right)_{\rm A} - \left(\frac{m_{\rm I}}{m_{\rm G}}\right)_{\rm B}\right] g'_{\rm h} l m_{\rm G}. \qquad (3.17)$$

In this way Eötvös greatly improved the limit of (3.14) to $\delta_{\rm AB} \leq 10^{-9}$. More recent experiments by others, ultimately involving the comparison of falling earth and moon in the solar gravitational field, have tightened this limit further to 1.5×10^{-13}.

3.2.2 EP and its significance

While preparing a review article on SR in 1907, Einstein came upon, what he later termed, "**my happiest thought**:" "Since all bodies accelerate the same way, an observer in a freely falling laboratory will not be able to detect any gravitational effect (on a point particle) in this frame." Or, "gravity is transformed away in reference frames in free fall."

Imagine an astronaut in a freely falling spaceship. Because all objects fall with the same acceleration, a released object in the spaceship will not be seen to fall. Thus, from the viewpoint of the astronaut, gravity is absent; everything becomes weightless.

To Einstein, this vanishing of the gravitational effect is so significant that he elevated it (in order to focus on it) to a physical principle: **the equivalence principle**.

$$\begin{pmatrix}\text{Physics in a frame freely falling in a gravity field}\\ \text{is equivalent to}\\ \text{Physics in an inertial frame without gravity}\end{pmatrix}.$$

Namely, within a freely falling frame, where the acceleration exactly cancels the uniform gravitational field, no sign of either acceleration or gravitation can be found by any physical means. Correspondingly,

$$\begin{pmatrix}\text{Physics in a nonaccelerating frame with gravity } \mathbf{g}\\ \text{is equivalent to}\\ \text{Physics in a frame without gravity but accelerating with } \mathbf{a} = -\mathbf{g}\end{pmatrix}.$$

Thus according to the EP of gravitation and inertia, accelerating frames of reference can be treated in exactly the same way as inertial frames. They are simply frames with gravity. From this we also obtain a **physics definition of an inertial frame**, without reference to any external environment such as fixed stars, as **the frame in which there is no gravity.** Einstein realized the unique position of gravitation in the theory of relativity. Namely, he understood that the question was not how to incorporate gravity into SR but rather how to use gravitation as a means to broaden the principle of relativity from inertial frames to all coordinate systems including the accelerating frames.

If we confine ourselves to the physics of mechanics, EP is just a re-statement of $m_{\mathrm{I}} = m_{\mathrm{G}}$. But once it is highlighted as a principle, it allowed Einstein to extend this equivalence to **all physics**: (not just to mechanics, but also electromagnetism, etc.) This generalized version is sometimes called the **strong equivalence principle**. Thus the "weak EP" is just the statement of $m_{\mathrm{I}} = m_{\mathrm{G}}$, while the "strong EP" is the principle of equivalence applied to all physics. In the following, we shall still call the strong equivalence principle as EP, for short.

Ordinarily we expect the gravitational effect to be very small as Newton's constant G_{N} is very small. One way to get an order of magnitude idea is by taking the ratio of the gravitational energy and the (relativistic) rest-energy[3] $E_{\mathrm{rel}} = m_{\mathrm{I}}c^2$

$$\psi = \frac{E_{\mathrm{grav}}}{E_{\mathrm{rel}}} = \frac{m_{\mathrm{G}}\Phi}{m_{\mathrm{I}}c^2} = \frac{\Phi}{c^2} = \frac{G_{\mathrm{N}}M}{c^2 r}, \tag{3.18}$$

where $\Phi = -G_{\mathrm{N}}M/r$ is the gravitational potential for the spherical symmetric case. Near Earth's surface, $\Phi = gh$, the product of gravitational acceleration and height, we have $\psi = gh/c^2 = O\,(10^{-15})$ for a typical laboratory distance range $h = O\,(10\text{ m})$ in the terrestrial gravity. Such a small value basically reflects the weakness of gravitational interaction. Only in extraordinary situations of black hole (extremely compact object) or cosmology (extreme massive system) with huge M to r ratios will the parameter ψ approach the order of unity. (For further discussion, see the introductory paragraphs of Chapter 7.)

[3] We assume that the reader being familiar with some basic relativistic physics such as $E = mc^2$, which will be properly derived in Chapter 10. See Eq. (10.37).

3.3 Implications of the strong EP

The strong EP implies, as we shall show in this section, that gravity can bend a light ray, shift the frequency of an electromagnetic wave, and cause clocks to run slow. Ultimately, these results suggested to Einstein that the proper framework to describe the relativistic gravitational effects is a curved spacetime.

To deduce the effect of gravity on certain physical phenomena, we shall use the following general procedure:

1. One first considers the description by an observer inside a spaceship in free fall. According to EP there is no gravitational effect in this inertial frame and SR applies.
2. One then considers the same situation from the viewpoint of an observer watching the spaceship from outside: there is a gravitational field and the first (freely falling) observer is seen to be accelerating in this gravitational field. Namely, this second observer defines an inertial frame, with gravity being treated as one of the forces $\mathbf{F} = m\mathbf{g}$.
3. The combined effects of acceleration and gravity, as seen by the second observer, must then reproduce the SR description as recorded by the inertial observer in free fall. "Physics should be independent of coordinate choices."

Bending of a light ray—a qualitative account

Let us first study the effect of gravity on a light ray traveling (horizontally) across a spaceship falling in a constant (vertical) gravitational field **g**. From the viewpoint of the astronaut in the spaceship, EP informs us, there is no detectable effect associated either with gravity or with acceleration: the light travels straight across the spaceship from one side to the other: in this coordinate frame, the light is emitted at a height h, received at the same height h on the opposite side of the spaceship, Fig. 3.3(a). But, to an observer outside the spaceship, there is a gravitational field **g** and the spaceship is accelerating (falling) in this gravitational field. The straight trajectory of the light signal in the freely falling spaceship will appear to bend, Fig. 3.3(b). Thus, to this outside observer, a light ray is seen to bend in the gravitational field.

We do not ordinarily see such falling of light rays because, for the gravitational field and distance scale that we are familiar with, this bending effect is unobservably small. Consider a lab with a width of 300 m. The duration for a light ray to travel across the lab would be 1 μs. During this interval, the distance y that the lab has fallen (amount of the bending) is extremely small: $y = gt^2/2 \simeq 5 \times 10^{-12}\,\text{m} = 0.05\,\text{Å}$. (Also cf. (3.18).) This EP consideration suggests that a light ray would be bent by any massive object. (The quantitative relation between the deflection angle and the gravitational potential will be worked out in Section 3.3.3.)

3.3.1 Gravitational redshift and time dilation

Gravitational redshift

In Fig. 3.3, we discussed the effect of a gravitational field on a light ray with its trajectory transverse to the field direction. Now let us consider the situation

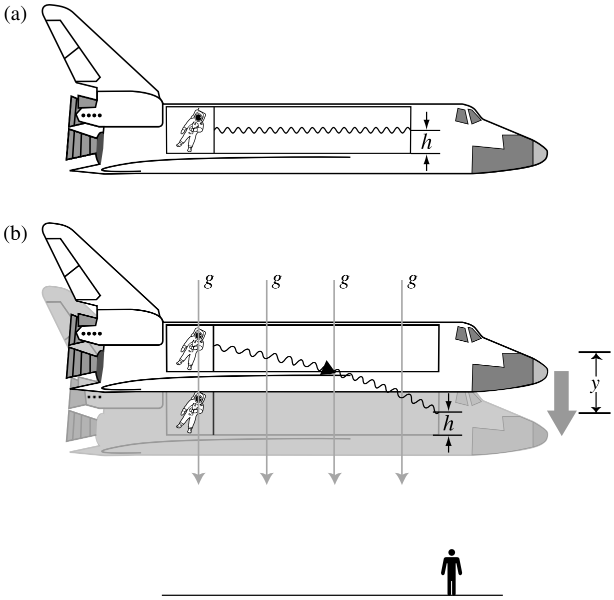

Fig. 3.3 According to EP, a light ray will "fall" in a gravitational field. (a) To the astronaut in the freely falling spaceship (an inertial observer), the light trajectory is straight. (b) To an observer outside the spaceship, the astronaut is accelerating (falling) in a gravitational field. The light ray will be bent so that it reaches the opposite side of the lab at a height $y = gt^2/2$ below the initial point.

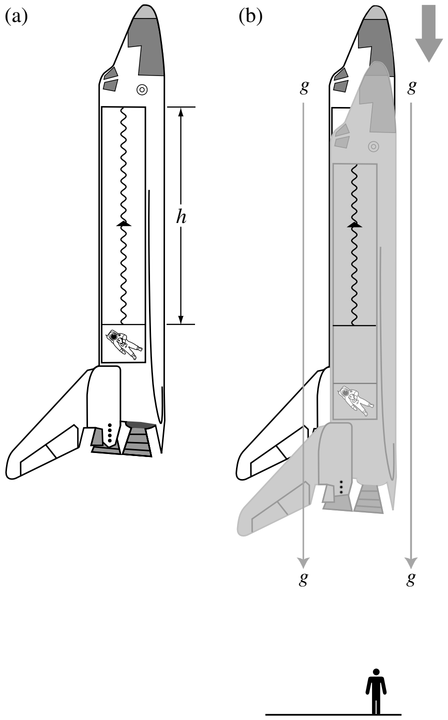

Fig. 3.4 According to EP, the frequency of a light ray is redshifted when moving up against gravity. (a) To an inertial observer in the freely falling spaceship, there is no frequency shift. (b) To an observer outside the spaceship, this astronaut is accelerating in a gravitational field, the null frequency shift result comes about because of the cancellation between a Doppler blueshift and a gravitational redshift.

when the field direction is parallel (or antiparallel) to the ray direction as in Fig. 3.4.

Here we have a receiver placed directly at a distance h above the emitter in a downward-pointing gravitational field $\mathbf{g}$. Just as the transverse case considered above, we first describe the situation from the viewpoint of the astronaut (in free fall), Fig. 3.4(a). EP informs us that the spaceship in free fall is an inertial frame. Such an observer will not be able to detect any physical effects associated with gravity or acceleration. In this free-fall situation, the astronaut should not detect any frequency shift: the received light frequency ω_{rec} is the same as the emitted frequency ω_{em}.

$$(\Delta\omega)_{\text{ff}} = (\omega_{\text{rec}} - \omega_{\text{em}})_{\text{ff}} = 0, \tag{3.19}$$

where the subscript ff reminds us that these are the values as seen by an observer in free fall.

From the viewpoint of the observer outside the spaceship, there is gravity and the spaceship is accelerating (falling) in this gravitational field, Fig. 3.4(b). Assume that this spaceship starts to fall at the moment of light emission. Because it takes a finite amount of time $\Delta t = h/c$ for the light signal to reach the receiver on the ceiling, it will be detected by a receiver in motion, with a velocity $\Delta u = g\Delta t$ (g being the gravitational acceleration). The familiar Doppler formula[4] (in the low-velocity approximation) would lead us to expect a frequency shift of

$$\left(\frac{\Delta\omega}{\omega}\right)_{\text{Doppler}} = \frac{\Delta u}{c}. \tag{3.20}$$

Since the receiver has moved closer to the emitter, the light waves must have been compressed, this shift must be toward the blue

$$(\Delta\omega)_{\text{Doppler}} = (\omega_{\text{rec}} - \omega_{\text{em}})_{\text{Doppler}} > 0. \tag{3.21}$$

We have already learned in (3.19), as deduced by the observer in free fall, that the received frequency did not deviate from the emitted frequency. This physical result must hold for both observers, this blueshift in (3.21) must somehow be cancelled. To the observer outside the spaceship, gravity is also present. We can recover the nullshift result, if there is another physical effect of **light being redshifted by gravity**, with just the right amount to cancel the Doppler blueshift of (3.20).

$$\left(\frac{\Delta\omega}{\omega}\right)_{\text{gravity}} = -\frac{\Delta u}{c}. \tag{3.22}$$

We now express the relative velocity on the RHS in terms of the gravitational potential difference $\Delta\Phi$ at the two locations

$$\Delta u = g\Delta t = \frac{gh}{c} = \frac{\Delta\Phi}{c}. \tag{3.23}$$

When (3.22) and (3.23) are combined, we obtain the phenomenon of **gravitational frequency shift**

$$\frac{\Delta\omega}{\omega} = -\frac{\Delta\Phi}{c^2}. \tag{3.24}$$

Namely,[5]

$$\frac{\omega_{\text{rec}} - \omega_{\text{em}}}{\omega_{\text{em}}} = -\frac{\Phi_{\text{rec}} - \Phi_{\text{em}}}{c^2}. \tag{3.25}$$

A light ray emitted at a lower gravitational potential point ($\Phi_{\text{em}} < \Phi_{\text{rec}}$) with a frequency ω_{em} will be received at a higher gravitational field point as a lower frequency ($\omega_{\text{em}} > \omega_{\text{rec}}$) signal, that is, it is redshifted, even though the emitter and the receiver are not in relative motion.

[4] The reader is assumed to know the Doppler effect. An abbreviated discussion is also provided in Chapter 10. See (10.48).

[5] Whether the denominator is ω_{rec} or ω_{em}, the difference is of higher order and can be ignored in these leading order formulae.

The Pound–Rebka–Snider experiment

In principle, this gravitational redshift can be tested by a careful examination of the spectral emission lines from an astronomical object (hence large gravitational potential difference). For a spherical body (mass M and radius R), the redshift formula of (3.24) takes on the form of

$$\frac{\Delta\omega}{\omega} = \frac{G_{\mathrm{N}}M}{c^2R}. \tag{3.26}$$

We have already commented on the smallness of this ratio in (3.18). Even the solar redshift has only a size $O\,(10^{-6})$, which can easily be masked by the standard Doppler shifts due to thermal motion of the emitting atoms. It was first pointed out by Eddington in the 1920s that the redshift effect would be larger for white dwarf stars, with their masses comparable to the solar mass and much smaller radii that the effect could be 10–100 times larger. But in order to obtain the mass measurement of the star, it would have to be in a binary configuration, for instance the Sirius A and B system. In such cases the light from the white dwarf Sirius B suffers scattering by the atmosphere of Sirius A. Nevertheless, some tentative positive confirmation of the EP prediction had been obtained. However, conclusive data did not exist in the first few decades after Einstein's paper. Surprisingly this EP effect of gravitational redshift was first verified in a series of terrestrial experiments when Pound and his collaborators (1960 and 1964) succeeded in measuring a truly small frequency shift of a radiation traveling up $h = 22.5$ m, the height of an elevator-shaft in the building housing the Harvard Physics Department:

$$\left|\frac{\Delta\omega}{\omega}\right| = \left|\frac{gh}{c^2}\right| = O\,(10^{-15}). \tag{3.27}$$

Normally, it is not possible to fix the frequency of an emitter or absorber to a very high accuracy because of the energy shift due to thermal recoils of the atoms. However, with the Mössbauer effect,[6] the emission line-width in a rigid crystal is as narrow as possible—limited only by the quantum mechanical uncertainty principle $\Delta t \Delta E \geq \hbar$, where Δt is given by the lifetime of the unstable (excited) state. Thus a long-lived state would have a particularly small energy-frequency spread. The emitting atom that Pound and Rebka chose to work with is an excited atom Fe*-57, which can be obtained through the nuclear beta-decay of cobalt-57. It makes the transition to the ground state by emitting a gamma ray: $\mathrm{Fe}^* \to \mathrm{Fe} + \gamma$. In the experiment, the γ-ray emitted at the bottom of the elevator shaft, after climbing the 22.5 m, could no longer be resonantly absorbed by a sheet of Fe in the ground state placed at the top of the shaft. To prove that the radiation has been redshifted by just the right amount $O\,(10^{-15})$, Pound and Rebka moved the detector slowly towards the emitter so that the (ordinary) Doppler blueshift is just the right amount to compensate for the gravitational redshift. In this way, the radiation is again absorbed. What was the speed with which they must move the receiver? From (3.24) and (3.20) we have

$$\frac{gh}{c^2} \underset{\text{gravity}}{=} \frac{\Delta\omega}{\omega} \underset{\text{Doppler}}{=} \frac{u}{c} \tag{3.28}$$

with

$$u = \frac{gh}{c} = \frac{9.8 \times 22.5}{3 \times 10^8} = 7.35 \times 10^{-7}\ \mathrm{m/s}. \tag{3.29}$$

[6] **The Mössbauer effect**—When emitting light, the recoil atom can reduce the energy of the emitted photon. In reality, since the emitting atom is surrounded by other atoms in thermal motion, this brings about recoil momenta in an uncontrollable way. (We can picture the atom as being part of a vibrating lattice.) As a result, the photon energy in different emission events can vary considerably, resulting in a significant spread of their frequencies. This makes it impossible for a measurement of the atomic frequency to high enough precision for purposes such as testing the gravitational redshift. But in 1958 Mössbauer made a breakthrough when he pointed it out, and verified by observation, that crystals with high Debye–Einstein temperature, that is, having a rigid crystalline structure, could pick up the recoil by the entire crystal. Namely, in such a situation, the emitting atom has an effective mass that is huge. Consequently, the atom loses no recoil energy, and the photon can pick up all the energy-change of the emitting atom, and the frequency of the emitted radiation is as precise as it can be.

It is such a small speed that it would take $h/u = c/g = O\ (3 \times 10^7\ \mathrm{s}) \simeq 1$ year to cover the same elevator shaft height. Of course this velocity is just the one attained by an object freely falling for a time interval that takes the light to traverse the distance h. This is the compensating effect we invoked in our derivation of the gravitational redshift at the beginning of this section.

Gravitational time dilation

At first sight, this gravitational frequency shift looks absurd. How can an observer, **stationary** with respect to the emitter, receive a different number of wave crests per unit time than the emitted rate? Here is Einstein's radical and yet simple answer: while the number of wave crests does not change, the time unit itself changes in the presence of gravity. The clocks run at different rates when situated at different gravitational field points: there is a **gravitational time dilation** effect.

Frequency being proportional to the inverse of local proper time rate

$$\omega \sim \frac{1}{d\tau} \tag{3.30}$$

the gravitational frequency shift formula (3.25) can be converted to a time dilation formula

$$\frac{d\tau_1 - d\tau_2}{d\tau_2} = \frac{\Phi_1 - \Phi_2}{c^2}, \tag{3.31}$$

or

$$d\tau_1 = \left(1 + \frac{\Phi_1 - \Phi_2}{c^2}\right) d\tau_2. \tag{3.32}$$

For static gravitational field, this can be integrated to read

$$\tau_1 = \left(1 + \frac{\Phi_1 - \Phi_2}{c^2}\right) \tau_2. \tag{3.33}$$

Namely, the clock at higher gravitational potential point will run faster. This is to be contrasted with the special relativistic time dilation effect—clocks in relative motion run at different rates. Here we are saying that two clocks, even at rest with respect to each other, also run at different rates if the gravitational fields at their respective locations are different. Their distinction can be seen in another way: in SR time dilation each observer sees the other's clock to run slow (see the subsection "relativity is truly relative" in Section A.1), while with gravitational dilation, the observer at a higher gravitational potential point sees the lower clock to run slow, and the lower observer sees the higher clock to run fast. For two clocks in a gravitational field and also in relative motion, we have to combine the gravitational and relative motion frequency-shift results to obtain

$$\tau_1 = \left(1 + 2\frac{\Delta\Phi}{c^2} - \frac{u^2}{c^2}\right)^{1/2} \tau_2. \tag{3.34}$$

Time dilation test by atomic clock The gravitational time dilation effects have been tested directly by comparing the times kept by two cesium atomic clocks: one flown in an airplane at high altitude h (about 10 km) in a holding pattern, for a long time τ, over the ground station where the other clock sits. After the correction of the various background effect (mainly SR time dilations), the high altitude clock was found to gain over the ground clock by a time interval of

$\Delta\tau = (gh/c^2)\tau$ in agreement with the expectation given in (3.33) (Hafele and Keating, 1972).

For an application in the Global Position System, see Problem 3.4.

A more direct derivation of time dilation

Instead of deriving the gravitational time dilation by way of the frequency shift result, we can obtain (3.33) more directly. Let us drop a clock in the gravitational field. It passes two locations at gravitational potential of Φ_1 and Φ_2, with velocities u_1 and u_2, respectively. The free fall frame being inertial (without gravity), the familiar SR formulae should apply. The time t^{ff} as recorded by this clock in free fall (the moving frame) to the times τ_1 and τ_2 as recorded by the clocks at these two locations (the rest frames) are related by the usual SR time dilation formulae:

$$t_1^{\mathrm{ff}} = \gamma_1\tau_1 \quad \text{and} \quad t_2^{\mathrm{ff}} = \gamma_2\tau_2 \tag{3.35}$$

with $\gamma_1 = 1/\sqrt{1-u_1^2/c^2}$. We are interested in the clock rates $d\tau_1$ and $d\tau_2$ given that $dt_1^{\mathrm{ff}} = dt_2^{\mathrm{ff}}$. The time dilation result of (3.32) can then be derived from (3.35):

$$\frac{d\tau_1}{d\tau_2} = \sqrt{\frac{1-u_1^2/c^2}{1-u_2^2/c^2}} \simeq 1 - \frac{1}{2}\frac{u_1^2-u_2^2}{c^2} = 1 + \frac{\Phi_1-\Phi_2}{c^2}, \tag{3.36}$$

where, at the second (approximate) equality, we have dropped $O(u^4/c^4)$ terms in the Taylor series expansions of the denominator and the square root. At the last equality we have used the low velocity version (consistent with our presentation) of the energy conservation relation $\frac{1}{2}m\Delta u^2 = -m\Delta\Phi$. Thus such gravitational time dilation effect is entirely compatible with the previously known SR time dilation effect—just as we have shown its compatibility with Doppler frequency shift in our first derivation (3.20) of gravitational redshift (3.24).

3.3.2 Light ray deflection calculated

The clocks run at different rates at locations where the gravitational field strengths are different. Since different clock rates will lead to different speed measurements, even the speed of light can be measured to have different values! We are familiar with light speed in different media being characterized by varying index of refraction. Gravitational time dilation implies that there is an effective index of refraction even in the vacuum when a gravitational field is present. Since gravitational field is usually inhomogeneous, this index is generally a position-dependent function.

Gravity-induced index of refraction in free space

At a given position r with gravitational potential $\Phi(r)$ a determination of the light speed involves the measurement of a displacement dr for a time interval $d\tau$ as recorded by a clock at rest at this position. The resultant ratio

$$\frac{dr}{d\tau} = c \tag{3.37}$$

is the light speed according to the local proper time. This speed c is a universal constant. Because of gravitational time dilation, as stated in (3.32), an observer at another position (with a different gravitational potential) would obtain a different value for this speed when using a clock located at the second position. In fact, a common choice of time coordinate is that given by a clock located far away from the gravitational source. For two positions $r_1 = r$ and $r_2 = \infty$, with r_2 being the reference point $\Phi(\infty) = 0$, while $\tau(r)$ is the local proper time, the clock at $r = \infty$ gives the coordinate time $t \equiv \tau(\infty)$. Equation (3.36) then yields the relation between the local time (τ) and the coordinate time (t) as

$$d\tau = \left(1 + \frac{\Phi(r)}{c^2}\right) dt. \tag{3.38}$$

This implies that the speed of light as measured by the remote observer is reduced by gravity as

$$c(r) \equiv \frac{dr}{dt} = \left(1 + \frac{\Phi(r)}{c^2}\right) \frac{dr}{d\tau} = \left(1 + \frac{\Phi(r)}{c^2}\right) c. \tag{3.39}$$

Namely, the speed of light will be seen by an observer (with his coordinate clock) to vary from position to position as the gravitational potential varies from position to position. For such an observer, the effect of the gravitational field can be viewed as introducing an **index of refraction** in the space:

$$n(r) \equiv \frac{c}{c(r)} = \left(1 + \frac{\Phi(r)}{c^2}\right)^{-1} \simeq 1 - \frac{\Phi(r)}{c^2}. \tag{3.40}$$

We will state the key concepts behind this position-dependent speed of light once more: we are not suggesting that the deviation of $c(r)$ from the constant c means that the physical velocity of light has changed, or that the velocity of light is no longer a universal constant in the presence of gravitational fields. Rather, it signifies that the clocks at different gravitational points run at different rates. For an observer, with the time t measured by clocks located far from the gravitational source (taken to be the coordinate time), the velocity of the light **appears to this observer** to slow down. A dramatic example is offered by the case of black holes (to be discussed in Section 6.4). There, as a manifestation of an infinite gravitational time dilation, it would take an infinite amount of coordinate time for a light signal to leave a black hole. Thus, to an outside observer, no light can escape from a black hole, even though the corresponding proper time duration is perfectly finite.

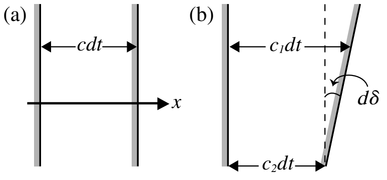

Fig. 3.5 Wavefronts of a light trajectory. (a) Wavefronts in the absence of gravity. (b) Tilting of wavefronts in a medium with an index of refraction varying in the vertical direction so that $c_1 > c_2$. The resultant light bending is signified by the small angular deflection $d\delta$.

Bending of light ray—the EP expectation

We can use this position-dependent index of refraction to calculate the bending of a light ray by a transverse gravitational field via the Huygen's construction. Consider a plane light wave propagating in the $+x$ direction. At each time interval Δt, a wavefront advances a distance of $c\Delta t$, see Fig. 3.5(a). The existence of a transverse gravitational field (in the y-direction) means a nonvanishing derivative of the gravitational potential $d\Phi/dy \neq 0$. Change of the gravitation potential means change in $c(r)$ and this leads to tilting of the wavefronts. We can then calculate the amount of the bending of the light ray by using the diagram in Fig. 3.5(b). A small angular deflection can be related to distances as

$$(d\delta) \backsimeq \frac{(c_1 - c_2)dt}{dy} \simeq \frac{d[c(r)](dx/c)}{dy}. \tag{3.41}$$

Working in the limit of weak gravity with small $\Phi(r)/c^2$ (or equivalently $n \simeq 1$), we can relate $d[c(r)]$ to a change of index of refraction as

$$d[c(r)] = cd[n^{-1}] = -cn^{-2}dn \simeq -cdn. \tag{3.42}$$

Namely, Eq. (3.14) becomes

$$(d\delta) \backsimeq -\frac{\partial n}{\partial y}dx. \tag{3.43}$$

But from (3.40) we have $dn(r) = -d\Phi(r)/c^2$, and thus, integrating (3.43), obtain the total deflection angle

$$\delta = \int d\delta = \frac{1}{c^2}\int_{-\infty}^{\infty} \frac{\partial \Phi}{\partial y}dx = \frac{1}{c^2}\int_{-\infty}^{\infty} (\nabla\Phi \cdot \hat{\mathbf{y}})dx. \tag{3.44}$$

The integrand is the gravitational acceleration perpendicular to the light path. We shall apply the above formula to the case of the spherical source $\Phi = -G_\mathrm{N}M/r$, and $\nabla\Phi = \hat{\mathbf{r}}G_\mathrm{N}M/r^2$. Although the gravitational field will no longer be a simple uniform field in the $\hat{\mathbf{y}}$ direction, our approximate result can still be used because the bending takes place mostly in the small region of $r \simeq r_\mathrm{min}$. See Fig. 3.6.

$$\delta = \frac{G_\mathrm{N}M}{c^2}\int_{-\infty}^{\infty} \frac{\hat{\mathbf{r}} \cdot \hat{\mathbf{y}}}{r^2}dx = \frac{G_\mathrm{N}M}{c^2}\int_{-\infty}^{\infty} \frac{y}{r^3}dx, \tag{3.45}$$

where we have used $\hat{\mathbf{r}} \cdot \hat{\mathbf{y}} = \cos\theta = y/r$. An inspection of Fig. 3.6 also shows that, for small deflection, we can approximate $y \backsimeq r_\mathrm{min}$, hence

$$r = (x^2 + y^2)^{1/2} \backsimeq (x^2 + r_\mathrm{min}^2)^{1/2} \tag{3.46}$$

leading to

$$\delta = \frac{G_\mathrm{N}M}{c^2}\int_{-\infty}^{\infty} \frac{r_\mathrm{min}}{(x^2 + r_\mathrm{min}^2)^{3/2}}dx = \frac{2G_\mathrm{N}M}{c^2 r_\mathrm{min}}. \tag{3.47}$$

With a light ray being deflected by an angle δ as shown in Fig. 3.6, the light source at S would appear to the observer at O to be located at S'. Since the deflection is inversely proportional to r_min, one wants to maximize the amount of bending by having the smallest possible r_min. For a light grazing the surface of the sun, $r_\mathrm{min} = R_\odot$ and $M = M_\odot$, Eq. (3.47) gives an angle of deflection $\delta = 0.875''$. As we shall explain in Section 6.2.1, this is exactly **half** of the correct GR prediction for the solar deflection of light from a distant star.

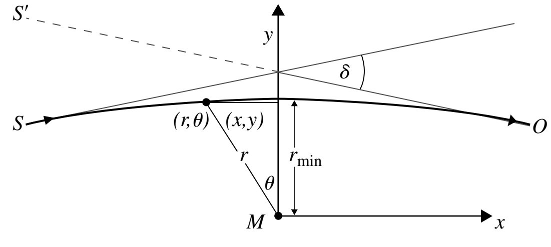

Fig. 3.6 Angle of deflection δ of light by mass M. A point on the light trajectory (solid curve) can be labeled either as (x, y) or (r, θ). The source at S would appear to the observer at O to be located at a shifted position of S'.

3.3.3 Energy considerations of a gravitating light pulse

Erroneous energy considerations

Because light gravitates (i.e. it bends and redshifts in a gravitational field), it is tempting to imagine that a photon has a (gravitational) mass. One might argue as follows: from the viewpoint of relativity, there is no fundamental difference between mass and energy, $E = m_{\mathrm{I}}c^2$. The equivalence $m_{\mathrm{I}} = m_{\mathrm{G}}$ means that any energy also has a nonzero "gravitational charge"

$$m_{\mathrm{G}} = \frac{E}{c^2}, \tag{3.48}$$

and hence will gravitate. The gravitational redshift formula (3.24) can be derived by regarding such a light-pulse losing "kinetic energy" when climbing out of a gravitational potential well. One can even derive the light deflection result (3.47) by using the Newtonian mechanics formula[7] of a moving mass (having velocity u) being gravitationally deflected by a spherically symmetric mass M (Fig. 3.6),

$$\delta = \frac{2G_{\mathrm{N}}M}{u^2 r_{\mathrm{min}}}. \tag{3.49}$$

For the case of the particle being a photon with $u = c$, this just reproduces (3.47). Nevertheless, such an approach to understand the effect of gravity on a light ray is **conceptually incorrect** because

- A photon is not a massive particle, and it cannot be described as a nonrelativistic massive object having a gravitational potential energy.
- This approach makes no connection to the underlying physics of gravitational time dilation.

The correct energy consideration

The energetics of gravitational redshift should be properly considered as follows (Schwinger, 1986; Okun *et al.*, 2000). Light is emitted and received through atomic transitions between two atomic energy levels of a given atom[8]: $E_1 - E_2 = \hbar\omega$. We can treat the emitting and receiving atoms as nonrelativistic massive objects. Thus when sitting at a higher gravitational potential point, the receiver atom gains energy with respect to the emitter atom,

$$E_{\mathrm{rec}} = E_{\mathrm{em}} + mgh.$$

We can replace the mass by (3.48) so that, to the leading order, $E_{\mathrm{rec}} = (1 + gh/c^2)E_{\mathrm{em}}$. One gets a multiplicative energy shift of the atomic levels. This implies that all the energy levels (and their differences) of the receiving atom are "blueshifted" with respect to those of the emitter atom by

$$(E_1 - E_2)_{\mathrm{rec}} = \left(1 + \frac{gh}{c^2}\right)(E_1 - E_2)_{\mathrm{em}}, \tag{3.50}$$

hence a fractional shift of atomic energy

$$\left(\frac{\Delta E}{E}\right)_{\mathrm{atom}} = \frac{gh}{c^2} = \frac{\Delta\Phi}{c^2}. \tag{3.51}$$

On the other hand, the traveling light pulse, neither gaining nor losing energy along its trajectory, has the **same** energy as the emitting atom. But it will be

[7] Equation (3.49) is quoted in small angle approximation of a general result that can be found in a textbook on mechanics. See, for example, Eq. (4.37) in (Kibble, 1985).

[8] We have used the fact that the energy of a light ray is proportional to its frequency. For most of us the quantum relation $E = \hbar\omega$ comes immediately to mind, but this proportionality also holds in classical electromagnetism where the field is pictured as a collection of harmonic oscillators.

seen by the blueshifted receiver atom as redshifted:

$$\left(\frac{\Delta E}{E}\right)_\gamma = -\frac{\Delta\Phi}{c^2} = \frac{\Delta\omega}{\omega}, \tag{3.52}$$

which is the previously obtained result (3.24). This approach is conceptually correct as

- Atoms can be treated as nonrelativistic objects having gravitational potential energy *mgh*.
- This derivation is entirely consistent with the gravitational time dilation viewpoint: the gravitational frequency shift does not result from any change of the photon property. It comes about because the standards of frequency (i.e. time) are different at different locations. This approach in fact gives us a physical picture of how clocks can run at different rates at different gravitational field points. An atom is the most basic form of a clock, with time rates being determined by transition frequencies. The fact that atoms have different gravitational potential energies (hence different energy levels) naturally give rise to different transitional frequencies, hence different clock rates.

The various results called "Newtonian"

The above discussion also explains why the usual erroneous derivations of treating photons as nonrelativistic massive particles with gravitational potential energy can lead to the correct EP formulae: the observed change of photon properties is due to the change in the standard clocks (atoms), which can be correctly treated as nonrelativistic masses with gravitational energy.

In this connection, we should also clarify the often-encountered practice of calling results such as (3.47) a Newtonian result. By this it is meant that the result can be derived in the pre-Einsteinian-relativity framework where particles can take on **any** speed we wish them to have. There does not exist the notion of a "low-velocity nonrelativistic limit." Consequently, it is entirely correct to use the mechanics formula (3.49) for a light particle which happens to propagate at the speed *c*.

However, one should be aware of the difference between this Newtonian (pre-relativistic) framework and the proper Newtonian limit, which we shall specify in later discussion, Sections 5.2.1 and 12.2.2, corresponding to the situation of nonrelativistic velocity, and static weak gravitational field. In this contemporary sense, (3.47) is not a result valid in the Newtonian limit.

3.3.4 Einstein's inference of a curved spacetime

Aside from the principle of relativity, EP is the most important physical principle underlying Einstein's formulation of a geometric theory of gravity. Not only allowing the accelerating frames to be treated on equal footing as the inertial frames and giving these early glimpses of the GR phenomenology, but also the study of EP physics led Einstein to propose that a curved spacetime is the gravitational field. We shall explain this connection in Chapter 5, after learning some mathematics of curved space in the following chapter.

Review questions

1. Write out, in terms of the gravitational potential $\Phi(x)$, the field equation and the equation of motion for Newton's theory of gravitation. What is the distinctive feature of this equation of motion (as opposed to that for other forces)?

2. What is the inertial mass? What is the gravitational mass? Give the simplest experimental evidence for their ratio being a universal constant (i.e. independent of material composition of the object).

3. What is the equivalence principle? What is weak EP? Strong EP?

4. Give a qualitative argument showing why EP can lead to the expectation of a gravitational bending of a light-ray.

5. Provide two derivations of the formula for gravitational frequency shift:
$$\frac{\Delta\omega}{\omega} = -\frac{\Delta\Phi}{c^2}.$$
(a) Use the idea that gravity can be transformed away by taking a reference frame in free fall; (b) Use the idea that atomic energy levels will be shifted in a gravitational field.

6. Derive gravitational time dilation formula
$$\frac{\Delta\tau}{\tau} = \frac{\Delta\Phi}{c^2}$$
in two ways: (a) from the gravitational frequency shift formula; (b) directly from the considerations of a clock in free fall.

7. Deduce the relation between coordinate time t (defined as the time measured by a clock located far away from any gravitational field) and local proper time $\tau(r)$ at a position with gravitational potential $\Phi(r)$:
$$dt = \frac{d\tau}{(1 + (\Phi/c^2))}.$$

8. The presence of a gravitational field implies the presence of an effective index of refraction in the free space. How does one arrive at this conclusion? Does this mean that speed of light is not absolute? Give an example of the physical manifestations of this index of refraction.

Problems

(3.1) **Inclined plane, pendulum, and EP:**

(a) **Inclined plane:** For the frictionless inclined plane (with angle θ) in Fig. 3.1(a), find acceleration's dependence on the ratio $m_{\mathrm{I}}/m_{\mathrm{G}}$. Thus a violation of the EP would show up as a material-dependence in the time for a material block to slide down the plane.

(b) **Pendulum:** For a simple pendulum on the surface of earth, cf. Fig. 3.1(b), find its oscillation period's dependence on the ratio $m_{\mathrm{I}}/m_{\mathrm{G}}$.

(3.2) **Two EP brain-teasers:**

(a) Use EP to explain the observation that a helium balloon **leans forward** in a (forward-) accelerating vehicle, see Fig. 3.7(a).

(b) On his 76th birthday Einstein received a gift from his Princeton neighbor Eric Rogers. It was a "toy" composed of a ball attached, by a spring, to the inside of a bowl, which was just the right size to hold the ball. The upright bowl is fastened to a broom-stick, see Fig. 3.7(b). What is the **surefire way**, as suggested by EP, to pop the ball back into the bowl each time?

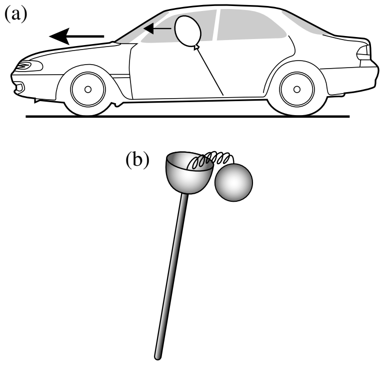

Fig. 3.7 Illustrations for the two EP brain-teasers in Problem 3.2.

(3.3) **Gravitational time dilation and the twin paradox** Consider the twin paradox given in Section A.1. Just before the traveling twin (Al) turned around his rocket-ship, his clock told him that the stay-at-home twin (Bill) had aged 9 years since his departure. But immediately after the turn-around, his clock found Bill's elapsed time to be 41 years. Use the gravitational time dilation effect to account for this change of 32 years.

(3.4) **The Global Position System** The signals from the 24 GPS satellites (in six evenly distributed orbit planes) enable us to fix our location on earth to a high degree of accuracy. Each satellite is at such an elevation so as to revolve around the earth every 12 h. In order to be accurate to within a few meters the satellite clocks must be highly accurate, as 10 ns time intervals translate into a light distance of 3 m. The atomic clocks on the satellites indeed have the capability of keeping time highly accurately, for example, to parts in 10^{13} over many days. (To be accurate over a long period, their time displays are remotely adjusted several times a day.) But in order to synchronize with the clocks on the ground for rapid determination of distances, we must take into account relativistic corrections. This calculation should make it clear that the proper functioning of the GPS requires our knowledge of relativity, especially GR. To investigate such relativistic effects we must first calculate the basic parameters of r_{s}, the satellite's radial distance (from the center of the earth), and v_{s}, its speed.

(a) Given the satellite orbit period being 12 h, calculate the speed v_{s} and distance r_{s}.

(b) Calculate the fractional change due to special relativistic time dilation.

(c) Calculate the fractional change due to the gravitational time dilation effect as the satellites are at a different gravitational potential compared to the surface of the earth. Is this GR effect more significant than the SR dilation?

(d) Calculate the error that can be accumulated in 1 min because of these relativistic corrections. Do these two effects change the satellite time in the same direction, or do they tend to cancel each other?

Metric description of a curved space

4

- Einstein's new theory of gravitation is formulated in a geometric framework of curved spacetime. In this chapter, we make a mathematical excursion into the subject of non-Euclidean geometry by way of Gauss's theory of curved surfaces.
- *Generalized (Gaussian) coordinates:* A systematical way to label points in space without reference to any objects outside this space.
- *Metric function:* For a given coordinate choice, the metric determines the intrinsic geometric properties of a curved space.
- *Geodesic equation:* It describes the shortest and the straightest possible curve in a warped space and is expressed in terms of the metric function.
- *Curvature:* It is a nonlinear second derivative of the metric. As the deviation from Euclidean relations is proportional to the curvature, it measures how much the space is warped.

By a deep study of the physics results implied by the equivalence principle (Chapter 3), Einstein proposed, as we shall discuss in the next chapter, that the gravitational field is the curved spacetime. Curved spacetime being the gravitational field, the proper mathematical framework for general relativity (GR) is Riemannian geometry and tensor calculus. We shall introduce these mathematical topics gradually. The key concepts are Gaussian coordinates, metric functions, and the curvature. Points in space are systematically labeled by the Gaussian coordinates; the geometry of space can be specified by length measurements among these points with results encoded in the metric function; the curvature tells us how much space is warped.

Historically Riemann's work on the foundation of geometry was based on an extension of Gauss's theory of curved surfaces. Since it is much easier to visualize two-dimensional (2D) surfaces, we shall introduce Riemannian geometry by first considering various topics in this case of the 2D curved space. In particular, we study the description of warped surfaces by a 2D metric g_{ab} with $a = 1, 2$, then suggest how such results can be generalized to higher n dimensions with $a = 1, 2, \ldots, n$.

Mathematically speaking, the algebraic extension to higher dimensional spaces, in particular the four-dimensional (4D) spacetime, is relatively straightforward—in most cases it involves an extension of the range of indices. Nevertheless, the generalization of the concept "curvature" to higher dimensions is highly nontrivial. A proper study of the curvature in higher dimensional spaces, called the Riemann curvature tensor, will be postponed until Chapter 11 when we present tensor analysis of GR. For the material in Parts I and II, we only need the metric description of the warped spacetime. Knowing the metric function, which is the relativistic gravitational potential, we can determine the equation of motion and can discuss many GR applications. In Part III we present the derivation of Riemann curvature tensor, and this finally allows us to write down Einstein equation, and to show that the metric functions used in Parts I and II are solutions to this GR field equation.

4.1 Gaussian coordinates

We first need an efficient way to label points in space. This section presents the study of the general coordinates in a curved space. For curved surfaces a position vector x^a has two independent components ($a = 1, 2$), and for higher dimensional cases, we concentrate particularly on the 4D spacetime $x^\mu (\mu = 0, 1, 2, 3)$.

Most of us start thinking of curved surfaces in terms of their embedding in the three-dimensional (3D) Euclidean space, in which the points can be labeled by Cartesian coordinate system (X, Y, Z). For illustration, we shall often use the example of a spherical surface, which geometers call a **2-sphere**.

A surface in the 3D space is specified by a **constraint condition**:

$$f(X, Y, Z) = 0, \tag{4.1}$$

or equivalently in the form of a relation among the coordinates,

$$Z = g(X, Y). \tag{4.2}$$

For the case of 2-sphere with radius R, such constraint conditions are

$$X^2 + Y^2 + Z^2 = R^2 \quad \text{or} \quad Z = \pm\sqrt{R^2 - X^2 - Y^2}. \tag{4.3}$$

This discussion of a curved surface being embedded in some larger space is an **extrinsic geometric description**—the physical space (here, the curved surface) is described using entities outside to this space. What we are most interested is an **intrinsic geometric description**—a characterization of the physical space without invoking any embedding. Namely, we are interested in the possibility of a description based solely on the measurement made by an inhabitant who never leaves the 2D surface. Gauss introduced a generalized parametrization, having coordinates (x^1, x^2) that are free to range over their respective domains without constraint.

$$X = X(x^1, x^2), \quad Y = Y(x^1, x^2), \quad Z = Z(x^1, x^2). \tag{4.4}$$

These generalized coordinates (x^1, x^2) are called the **Gaussian coordinates.** We note the employment of Gaussian coordinates avoids such constraint expressions as in (4.2). Using the Gaussian coordinates (their number being the dimensionality of the space) the geometric description can be purely intrinsic.

For the case of 2-sphere (see Fig. 4.1), we illustrate Gaussian coordinate choices by two systems:

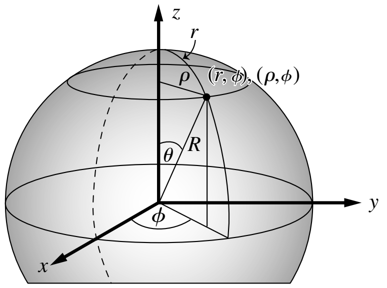

Fig. 4.1 Gaussian coordinates (r, ϕ) and (ρ, ϕ) for the curved surface of a 2-sphere. The dashed line is the prime meridian. NB, the radial coordinate r is taken in reference to the "north pole" on the surface of the sphere, rather with respect to the center of the sphere.

1. **The polar coordinate system**: We can set up a Gaussian coordinate system $(x^1, x^2) = (r, \phi)$ by first picking a point on the surface to be the origin (the "north pole") and a longitudinal great circle as the "prime

meridian." The radial coordinate r is marked on the spherical surface in the radial direction and the azimuthal angle ϕ is measured against the prime meridian. Thus, r has the range of $0 \leq r \leq \pi R$ and is related to the polar angle[1] as $r = R\theta$ where $0 \leq \theta \leq \pi$ and $0 \leq \phi \leq 2\pi$.

$$X = R\sin\frac{r}{R}\cos\phi, \quad Y = R\sin\frac{r}{R}\sin\phi, \quad Z = R\cos\frac{r}{R}. \tag{4.5}$$

2. **The cylindrical coordinate system**: We can choose another set of Gaussian coordinates by having a different radial coordinate: instead of r, we pick the function $\rho = R\sin\theta = R\sin(r/R)$ as our radial coordinate with a range of $0 \leq \rho \leq R$. Namely, we now have the system $(x^1, x^2) = (\rho, \phi)$. If the spherical surface is embedded in a 3D Euclidean space, ρ is interpreted as the perpendicular distance to the z-axis as shown in Fig. 4.1.

$$X = \rho\cos\phi, \quad Y = \rho\sin\phi, \quad Z = \pm\sqrt{R^2 - \rho^2}. \tag{4.6}$$

From now on we will no longer use the extrinsic coordinates such as (X, Y, Z). By coordinates, we shall always mean the Gaussian coordinates (x^1, x^2) as the way to label points on a 2D space. Since one could have chosen any number of coordinate systems, and at the same time expecting geometric relations to be independent of such choices, a proper formulation of geometry must be such that it is invariant under general coordinate transformations.

4.2 Metric tensor

The central idea of differential geometry was that an intrinsic description of space could be accomplished by distance measurements made within physical space. Namely, one can imagine labeling various points of space (with a Gaussian coordinate system), then measure the distance among neighboring points. From the resultant "table of distance measurements," one obtains a description of this space. For a given coordinate system, these measurements are encoded in the metric function.

In fact, we have already used this Gaussian prescription in Chapter 2 when we first introduced the notion of a metric directly in terms of the basis vectors $\{\mathbf{e}_a\}$ defined within the physical space (cf. (2.33)):

$$g_{ab} = \mathbf{e}_a \cdot \mathbf{e}_b. \tag{4.7}$$

The metric g_{ab} relates the (infinitesimal) length measurement ds to the chosen coordinates $\{dx^a\}$, as shown in (2.36):

$$\begin{aligned} ds^2 &= g_{ab}dx^a dx^b && (4.8)\\ &= g_{11}(dx^1)^2 + g_{22}(dx^2)^2 + 2g_{12}(dx^1 dx^2), && (4.9)\end{aligned}$$

where in (4.8) Einstein's convention of summing over repeated indices has been employed. We have also used the symmetry property of the metric, $g_{12} = g_{21}$. The above relation can also be written as a matrix equation,

$$ds^2 = (dx^1 \quad dx^2)\begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix}\begin{pmatrix} dx^1 \\ dx^2 \end{pmatrix} \tag{4.10}$$

with the metric being represented by a 2 × 2 matrix.

[1] We can of course pick (θ, ϕ) as our Gaussian coordinates. But this is essentially the same system as the polar coordinates being discussed here.

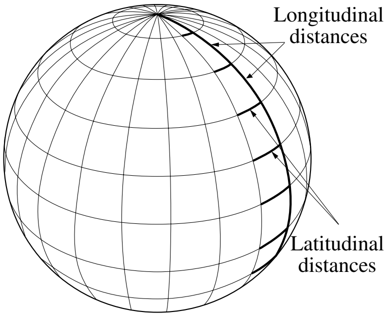

Fig. 4.2 Using distance measurements along longitudes and latitudes to specify the shape of the spherical surface.

Metric in polar coordinates To illustrate this for the case of a spherical surface (radius R), one first sets up the latitude/longitude system (i.e. a system of polar coordinates r and ϕ) to label points on the globe, then measures the distances between neighboring points (Fig. 4.2). One finds that the latitudinal distances ds_ϕ (subtended by $d\phi$ between two points having the same r value) become ever smaller as one approaches the poles $ds_\phi = R\sin\theta d\phi = R\sin(r/R)d\phi$, while the longitudinal distance interval dr between two points at the same longitude ($d\phi = 0$) can be chosen to have the same value over the whole range of θ and ϕ. From such a table of distance measurements, one obtains a description of this spherical surface. Such distance measurements can be compactly expressed in terms of the metric tensor elements. Because we have chosen an orthogonal coordinate $g_{r\phi} = \mathbf{e}_r \cdot \mathbf{e}_\phi = 0$,

$$[ds^2]_{(r,\phi)} = (ds_r)^2 + (ds_\phi)^2 \tag{4.11}$$

$$= dr^2 + R^2 \sin^2 \frac{r}{R} d\phi^2. \tag{4.12}$$

NB an infinitesimally small area on a curved surface can be thought of as a (infinitesimally small) flat plane. For such a flat surface ds can be calculated by Pythagorean theorem as in (4.11). A comparison of (4.12) and (4.10) leads to an expression for the metric tensor for this coordinate system to be

$$g_{ab}^{(r,\phi)} = \begin{pmatrix} 1 & 0 \\ 0 & R^2\sin^2(r/R) \end{pmatrix}. \tag{4.13}$$

Metric in cylindrical coordinates To calculate the metric for spherical surface with cylindrical Gaussian coordinates (ρ, ϕ) as shown in (4.6). From Fig. 4.1 we see that the cylindrical radial coordinate ρ is related to the polar angle $\theta = r/R$ by $\rho = R\sin(r/R)$, hence $d\rho = \sqrt{1 - (\rho^2/R^2)}dr$. From this and from (4.12) we obtain

$$[ds^2]_{(\rho,\phi)} = \frac{R^2 d\rho^2}{R^2 - \rho^2} + \rho^2 d\phi^2, \tag{4.14}$$

corresponding to the metric

$$g_{ab}^{(\rho,\phi)} = \begin{pmatrix} R^2/(R^2-\rho^2) & 0 \\ 0 & \rho^2 \end{pmatrix}. \tag{4.15}$$

We are interested in the cylindrical coordinate system also because this offers, as we shall show in Section 4.3.2, a rather compact description of all curved surfaces with constant curvature.

We emphasize it again: the metric g_{ab} is an **intrinsic geometric quantity** because it can be determined without reference to any embedding—a 2D inhabitant on the curved surface can, once the Gaussian coordinates $\{x^a\}$ have been chosen, obtain g_{ab} by various length ds-measurements spanned by dx^a and dx^b. For the 2D case, we have

$$g_{11} = \frac{(ds_1)^2}{(dx^1)^2}, \quad g_{22} = \frac{(ds_2)^2}{(dx^2)^2}, \tag{4.16}$$

$$g_{12} = \frac{(ds_{12})^2 - (ds_1)^2 - (ds_2)^2}{2dx^1 dx^2}, \tag{4.17}$$

where ds_1 and ds_2 are the lengths measured along the 1- and 2-axes having respective coordinates dx^1 and dx^2, and ds_{12} the length of a segment on the 1-2 plane (with coordinate projections dx^1 and dx^2 along the axes). From the law of cosines,[2] we see that (4.17) just says that g_{12} is the cosine of the angle subtended by the axes, cf. (4.7) and (4.10). Thus if we had an orthogonal coordinate system the metric matrix would be diagonal. It should be emphasized that coordinates $\{x^a\}$ themselves do **not** measure distance. Only through the metric as in (4.8) are they connected to distance measurements.

General coordinate transformation

For a description of the spherical surface one can use, for example, either polar Gaussian coordinates: $(x^1, x^2) = (r, \phi)$, or cylindrical Gaussian coordinates $(x'^1, x'^2) = (\rho, \phi)$. The (intrinsic) geometric properties of the spherical surface are of course independent of the coordinate choice. For example, the length should be unchanged $ds^2 = g_{ab}dx^a dx^b = g'_{ab}dx'^a dx'^b$. Here, these two coordinate systems are related by $\rho = R\sin(r/R)$, or $d\rho = \cos(r/R)dr$. This change of coordinates $(r, \phi) \to (\rho, \phi)$ can be expressed as a transformation matrix acting on the coordinate differentials that leaves the infinitesimal length intervals ds^2 unchanged.

$$\begin{pmatrix} d\rho \\ d\phi \end{pmatrix} = \begin{pmatrix} \cos(r/R) & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} dr \\ d\phi \end{pmatrix}. \tag{4.18}$$

Equation (4.18) can be compared to the similar coordinate transformations of rotation and boost that we discussed previously, for example, in Section 2.3.2. We note an important difference: the elements of the transformation matrices in (2.54) and (2.60) are not position-dependent—the same rotation angle θ and the same boost velocity v for every point in space. The matrix in (4.18) has position-dependent elements. Namely, we make a different transformation (from the polar to the cylindrical system) at each position having a different radial coordinate r. This is a key difference between the coordinate transformation in a flat space and those in curved space. To have coordinate changes being dependent on coordinates themselves means that the transformation is nonlinear. This will be discussed extensively in Part III (Chapters 10 and 11) when we present the tensor calculus in SR (flat spacetime) vs. that in GR (curved spacetime).

4.2.1 Geodesic as the shortest curve

So far in our introductory discussion of the metric we have used known curved surfaces such as sphere to show that its shape can be specified by the metric function. This justifies our subsequent application that a curved space can be represented, for a given coordinate system, by a metric tensor. Once we have the metric, other geometric quantities can then be computed. For example, angles can be determined:

$$\cos\theta = \frac{\mathbf{A}\cdot\mathbf{B}}{AB} = \frac{g_{ab}A^aB^b}{\sqrt{g_{cd}A^cA^d}\sqrt{g_{ef}B^eB^f}}. \tag{4.19}$$

Our main task here is to show that the curve having the extremum length, called the **geodesic line**, can be specified in terms of the metric function.

[2] The flat space geometric relation is applicable because we are working in an infinitesimal region, which can always be described approximately as a flat space, cf. Section 4.2.2.

Any curve can be represented by a set of coordinates $x^a(\sigma)$ depending on a single parameter σ, which has some definite range.[3] In a curved space the metric only determines the infinitesimal length:

$$ds = \sqrt{g_{ab}dx^a dx^b}. \tag{4.20}$$

For a finite length, we must perform the line-integration,

$$s = \int ds = \int \frac{ds}{d\sigma}d\sigma = \int \sqrt{\left(\frac{ds}{d\sigma}\right)^2} d\sigma = \int L\, d\sigma, \tag{4.21}$$

where L is the "Lagrangian":

$$L = \sqrt{g_{ab}\frac{dx^a}{d\sigma}\frac{dx^b}{d\sigma}} = L(x, \dot{x}) \tag{4.22}$$

with $\dot{x} = dx/d\sigma$. To determine the shortest (i.e. the extremum) line in the curved space, we impose the extremization condition for variation of the path with end points fixed:

$$\delta s = \delta \int L(x, \dot{x})d\sigma = 0, \tag{4.23}$$

which can be translated, by calculus of variation, into a partial differential equation—the Euler–Lagrange equation:

$$\frac{d}{d\sigma}\frac{\partial L}{\partial \dot{x}^a} - \frac{\partial L}{\partial x^a} = 0. \tag{4.24}$$

To aid the reader in recalling this connection, which is usually learnt in an intermediate mechanics course, we provide a brief derivation for the simple one-dimensional (1D) case. We set out to minimize the 1D integral s with respect to the variation, not of one variable or several variables as in the usual minimization problem, but of a whole function $x(\sigma)$ with initial and final values fixed:

$$\delta x(\sigma) \quad \text{with } \delta x(\sigma_i) = \delta x(\sigma_f) = 0. \tag{4.25}$$

The variation of the integrand being

$$\delta L(x, \dot{x}) = \frac{\partial L}{\partial x}\delta x + \frac{\partial L}{\partial \dot{x}}\delta \dot{x}, \tag{4.26}$$

we have

$$\begin{aligned} 0 = \delta s = \delta \int_{\sigma_i}^{\sigma_f} L(x, \dot{x})d\sigma &= \int_{\sigma_i}^{\sigma_f} \left(\frac{\partial L}{\partial x}\delta x + \frac{\partial L}{\partial \dot{x}}\frac{d}{d\sigma}\delta x\right) d\sigma \\ &= \int_{\sigma_i}^{\sigma_f} \left(\frac{\partial L}{\partial x} - \frac{d}{d\sigma}\frac{\partial L}{\partial \dot{x}}\right)(\delta x)d\sigma. \end{aligned} \tag{4.27}$$

To reach the last expression we have performed an integration-by-parts on the second term, and used the condition in (4.25) to discard the integrated term $[(\partial L/\partial \dot{x})\delta x]_{\sigma_i}^{\sigma_f}$. Since δs must vanish for arbitrary variations $\delta x(\sigma)$, the expression in the parenthesis must vanish. This is the one-dimension version of the Euler–Lagrange Eq. (4.24). In mechanics, the curve parameter is time $\sigma = t$ and Lagrangian L is simply the difference between kinetic and potential energy. For the simplest case of $L = \frac{1}{2}m\dot{x}^2 - V(x)$, the Euler–Lagrange equation is just the familiar $F = ma$ equation.

[3] For a concrete example, one can think of the trajectory of a particle as a curve $\mathbf{x}(t)$ with the curve parameter being the time variable.

As a mathematical exercise, one can show that the **same** Euler–Lagrange Eq. (4.24) follows from, instead of (4.22), a Lagrangian of the form:

$$L(x, \dot{x}) = g_{ab}\dot{x}^a\dot{x}^b, \tag{4.28}$$

which without the square-root is much easier to work with than (4.22). With L in this form, the derivatives become

$$\frac{\partial L}{\partial \dot{x}^a} = 2g_{ab}\dot{x}^b, \quad \frac{\partial L}{\partial x^a} = \frac{\partial g_{cd}}{\partial x^a}\dot{x}^c\dot{x}^d, \tag{4.29}$$

where we have used the fact that the metric function g_{ab} depends on x, but not $\dot{x}$. Substituting these relations back into Eq. (4.24), we obtain the **geodesic equation**,

$$\frac{d}{d\sigma}g_{ab}\dot{x}^b - \frac{1}{2}\frac{\partial g_{cd}}{\partial x^a}\dot{x}^c\dot{x}^d = 0, \tag{4.30}$$

which determines the trajectory of the "shortest curve." One can easily use this equation to check the geodesic lines in simple surfaces of flat plane and spherical surface (Problem 4.4).

4.2.2 Local Euclidean coordinates

We are familiar with the idea that at any point on a curved surface there exists a plane, tangent to the curved surface. The plane in its Cartesian coordinates, can have a metric δ_{ab}. But this is true only at this point (call it the origin). Namely, $\bar{g}_{ab}(0) = \delta_{ab}$. If we are interested in the metric function, we have to be more careful. A more complete statement is given by the flatness theorem.

The flatness theorem: In a curved space with a general coordinate system x^a and a metric value g_{ab} at a given point P, we can always find a coordinate transformation $x^a \to \bar{x}^a$ and $g_{ab} \to \bar{g}_{ab}$ so that the metric is flat at this point: $\bar{g}_{ab} = \delta_{ab}$ and $\partial\bar{g}_{ab}/\partial\bar{x}^c = 0$,

$$\bar{g}_{ab}(\bar{x}) = \delta_{ab} + \gamma_{abcd}(0)\bar{x}_c\bar{x}_d + \cdots. \tag{4.31}$$

Namely the metric in the neighborhood of the origin will differ from δ_{ab} by the second order derivative. This is simply a Taylor series expansion of the metric at the origin—there is the constant $\bar{g}_{ab}(0)$ plus higher order derivative terms ($\gamma_{abcd}(0)\bar{x}_c\bar{x}_d$ being simply the second derivative). The nontrivial content of (4.31) is the absence of the first derivative. That $\bar{g}_{ab}(0) = \delta_{ab}$ should be less surprising: it is not difficult to see that for a metric value at one point one can always find an orthogonal system so that $\bar{g}_{ab}(0) = 0$ for $a \neq b$ and the diagonal elements can be scaled to unity so that the new coordinate bases all have unit length and the metric being an identity matrix. If the original metric has negative determinant, this reduces to the pseudo-Euclidean metric $\eta_{ab} = \text{diag}(1, -1)$ (cf. Problem 4.5). Such a coordinate system $\{\bar{x}^a\}$ is called the locally Euclidean frame (LEF).

The theorem can be generalized to n-dimensional space, in particular the 4D spacetime. (The proof will be provided in Box 11.1.) It informs us that the general spacetime metric $g_{ab}(x)$ is characterized at a point (P) not so much by the value $g_{ab}|_P$ since that can always be chosen to be flat, $\bar{g}_{ab}|_P = \delta_{ab}$, nor by its first derivative which can always be chosen to vanish $\partial\bar{g}_{ab}/\partial x^c|_P = 0$, but by the second derivative of the metric $\partial^2 g_{ab}/\partial x^c \partial x^d$, which are related to the curvature to be discussed in Section 4.3.

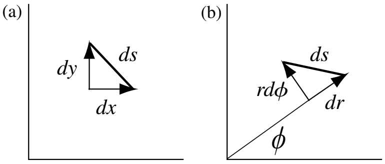

Fig. 4.3 Two coordinate systems in a flat plane: (a) Cartesian coordinates, and (b) polar coordinates.

Box 4.1 More illustrative calculations of metric tensors for simple surfaces

Here are further examples of metric tensors for 2D surfaces, calculated by using the fact that any surface in the small can be approximated by a plane having Cartesian coordinates. We also take this occasion to discuss the possibility of using the metric tensor to determine whether a surface is curved or not. For a flat surface, we can find a set coordinate so that the metric tensor is position-independent; this cannot be done in a curved surface as $g_{ab} = \mathbf{e}_a \cdot \mathbf{e}_b$ must change from point to point.

1. **A plane surface with Cartesian coordinates**: For the coordinates $(x^1, x^2) = (x, y)$, we have the infinitesimal length $ds^2 = dx^2 + dy^2$, Fig. 4.3(a). Comparing this to the general expression in (4.8), we see that the metric must be

$$g_{ab} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{4.32}$$

which is, of course, position-independent. This is possible only if the space is not curved.

2. **A plane surface with polar coordinates**: For the coordinates $(x^1, x^2) = (r, \phi)$, we have the infinitesimal length $ds^2 = (dr)^2 + (rd\phi)^2$, Fig. 4.3(b), thus according to (4.8), a metric

$$g_{ab} = \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix}, \tag{4.33}$$

which is position-dependent! But we can find a coordinate transformation $(x^1, x^2) \to (x'^1, x'^2)$ so that the metric in the new coordinate is position-independent, $g'_{ab} = \delta_{ab}$. Of course, the new coordinates are just the Cartesian coordinates $(x'^1, x'^2) = (x, y)$:

$$x = r\cos\phi, \quad y = r\sin\phi. \tag{4.34}$$

3. **A cylindrical surface with cylinder coordinates**: Let R be the radius of the cylinder (see Fig. 4.4(a)), the infinitesimal length for cylinder coordinates $(x^1, x^2) = (z, R\phi)$ is then $ds^2 = dz^2 + R^2 d\phi^2 = (dx^1)^2 + (dx^2)^2$. This shows that we have a constant metric $g_{ab} = \delta_{ab}$. Thus locally this is a flat surface, even though globally and topologically it is different from a plane surface. For example, a straight line can close onto itself in such a cylindrical surface (see Fig. 4.4(b)).

4. **A spherical surface with spherical coordinates**: For a spherical surface with radius R, we have already calculated the metric: for the polar coordinates (r, ϕ) in (4.13) and for cylindrical coordinates (ρ, ϕ) in (4.15). They are all position-dependent. Furthermore, such position dependence cannot be transformed away by going to any other coordinate system.

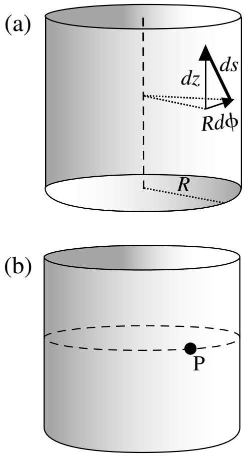

Fig. 4.4 (a) Cylindrical coordinates on a cylindrical surface. (b) A straight line on a cylindrical surface can return to the originating point.

4.3 Curvature

From the above discussion, we see that the metric value cannot represent the essence of a curved space because it is coordinate-dependent and can always, at a given point, be transformed to a flat space metric. However, this replacement can only be done locally—in an infinitesimally small region. A general metric equates to the flat space metric up to corrections given by the second derivative of the metric function, as shown in (4.31). This suggests that it is the second derivatives that really tell us how curved a curved space is.

4.3.1 Gaussian curvature

Consider the three surfaces in Fig. 4.5. We usually can tell whether a surface is curved by an examination of its relation to the embedding space. Thus the sphere is curved in a fundamental way. By contrast, the curvature of the cylinder is less fundamental as we can cut and unroll it into a plane without internal deformation—we say it has zero intrinsic curvature (although such a cylinder has global curvature), cf. Box 4.1, item 3. We are interested in finding a simple **intrinsic** method to determine whether a space is warped or not.

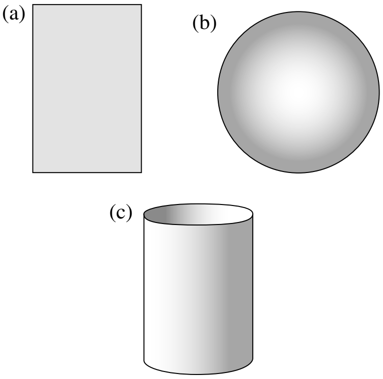

Fig. 4.5 Three kinds of surfaces: (a) flat plane, (b) sphere, and (c) cylindrical surface.

For a flat space we can find a coordinate system such that the metric is position-independent, while such coordinates do not exist in the case of a curved space. Using the metric in this way to determine whether a surface is curved or not, as in Box 4.1, is rather unsatisfactory. How can we be sure to have exhausted all possible coordinate systems, none of which have a constant metric? Is there a better way?

Theorema Egregium (a very beautiful theorem): This is the title of the paper in which Gauss presented his answer to the above questions: he showed that it was possible to define an unique invariant second derivative of the metric tensor ($\partial^2 g$) called the **curvature** K, such that, independent of the coordinate choice, $K = 0$ for a flat and $K \neq 0$ for curved surfaces.

With no loss of generality we shall quote Gauss's result for a diagonalized metric $g_{ab} = \text{diag}\,(g_{11}, g_{22})$:

$$K = \frac{1}{2g_{11}g_{22}}\left\{-\frac{\partial^2 g_{11}}{(\partial x^2)^2} - \frac{\partial^2 g_{22}}{(\partial x^1)^2} + \frac{1}{2g_{11}}\left[\frac{\partial g_{11}}{\partial x^1}\frac{\partial g_{22}}{\partial x^1} + \left(\frac{\partial g_{11}}{\partial x^2}\right)^2\right]\right.$$
$$\left. + \frac{1}{2g_{22}}\left[\frac{\partial g_{11}}{\partial x^2}\frac{\partial g_{22}}{\partial x^2} + \left(\frac{\partial g_{22}}{\partial x^1}\right)^2\right]\right\}. \tag{4.35}$$

Since this curvature is expressed entirely in terms of the metric and its derivatives, it is also an intrinsic geometric object. This one quantity determines the curvature of a surface—in contrast to the embedded viewpoint, which may lead one to expect that it would take two numbers to characterize the curvature of a 2D space. In fact, there is no curvature for an 1D space; an inhabitant on a line cannot detect any intrinsic curvature. To describe such curvature of a 2D surface, it only takes one number. We will not present the derivation of (4.35) since it is contained in the more general result to be discussed in Section 11.3.1

(Problem 11.11). But, let us check that it indeed has the property as a simple indicator as whether a surface is warped or not.

- For a position-independent metric we automatically have $K = 0$ because the derivatives of the metric vanish. Thus for the plane surface with Cartesian coordinates and the cylindrical surface with cylindrical coordinates we can immediately conclude that they are intrinsically flat surfaces.
- For a plane surface with polar coordinates $(x^1, x^2) = (r, \phi)$, we have a position-dependent metric $g_{11} = 1$ and $g_{22} = r^2 = (x^1)^2$ with $(\partial g_{22}/\partial x^1) = 2x^1$ and $(\partial^2 g_{22}/\partial(x^1)^2) = 2$. However, the curvature vanishes:
$$K = \frac{1}{2(x^1)^2}\left\{-2 + \frac{1}{2(x^1)^2}\left[4(x^1)^2\right]\right\} = 0 \tag{4.36}$$
indicating that it is a flat space, even though the corresponding metric (4.33) is position-dependent.
- For a spherical surface with polar coordinates $(x^1, x^2) = (r, \phi)$, we have Eq. (4.13) having $g_{11} = 1$ and $g_{22} = R^2 \sin^2(x^1/R)$ with $\partial g_{22}/\partial x^1 = R\sin(2x^1/R)$ and $\partial^2 g_{22}/(\partial x^1)^2 = 2\cos(2x^1/R)$. This leads to
$$\begin{aligned} K = &\frac{1}{2R^2 \sin^2(x^1/R)}\left\{2\sin^2\frac{x^1}{R} - 2\cos^2\frac{x^1}{R}\right. \\ &\left. + \frac{4R^2\sin^2(x^1/R)\cos^2(x^1/R)}{2R^2\sin^2(x^1/R)}\right\} = \frac{1}{R^2}. \end{aligned} \tag{4.37}$$
One can easily check that this result holds for the cylindrical coordinates of (4.15) as well, indicating that $K = R^{-2}$ is the curvature for a spherical surface, independent of coordinate choices.

4.3.2 Spaces with constant curvature

In Chapter 7 we shall start our discussion of cosmology with the basic assumption (called the cosmological principle) that the space, at given instant of cosmic time, is homogeneous and isotropic. Not surprisingly, this corresponds to a 3D space of constant curvature. Since it is difficult to visualize a warped 3D space, we shall first discuss the 2D surface with constant curvature. The generalization of this result to 3D space will then be presented afterward.

2D surfaces with constant curvature

While the Gaussian curvature in (4.35) is generally a position-dependent function, we have seen in (4.37) that the sphere has a constant $K = 1/R^2$. Obviously, a flat plane is a surface of constant curvature, $K = 0$. In fact there are three surfaces having constant curvatures:
$$K = \frac{k}{R^2}, \tag{4.38}$$
with the **curvature signature** $k = +1, 0$, and -1. Namely, besides the two familiar surfaces of 2-sphere and flat plane, there is another surface, called a **2-pseudosphere**, with a negative curvature $K = -1/R^2$.

What should be the metric for pseudosphere so that (4.35) can yield a negative curvature? An inspection of the calculation in (4.37) shows that, in order to obtain a result of $-1/R^2$, we would want the first term in the curly parenthesis to change sign (since the next two terms cancel each other). This first term originates from $\cos(2x^1/R) = -\sin^2(x^1/R)+\cos^2(x^1/R)$ in the second derivative of g_{22}. This suggests that, to go from the positive curvature for a sphere to a negative curvature for a pseudosphere, the metric term $g_{22} = R^2 \sin^2(x^1/R)$ should be changed to $g_{22} = R^2 \sinh^2(x^1/R)$ so that the second derivative of the new g_{22} would have a factor of $\cosh(2x^1/R) = +\sinh^2(x^1/R) + \cosh^2(x^1/R)$. Making such a change in Eq. (4.13), we have the metric for the pseudosphere

$$g_{ab}^{(r,\phi)} = \begin{pmatrix} 1 & 0 \\ 0 & \sin h^2(r/R) \end{pmatrix}. \tag{4.39a}$$

which leads to the curvature:

$$\begin{aligned} K =& \frac{1}{2R^2\sinh^2(x^1/R)}\Bigg\{-2\sinh^2\frac{x^1}{R} - 2\cosh^2\frac{x^1}{R} \\ &+ \frac{4R^2\sinh^2(x^1/R)\cosh^2(x^1/R)}{2R^2\sinh^2(x^1/R)}\Bigg\} = \frac{-1}{R^2}. \end{aligned} \tag{4.39}$$

Such a negative curvature space is also referred to as a hyperbolic space.

In this way the infinitesimal separation for the three surfaces with constant curvature in the polar coordinates as shown in (4.13), (4.33), and (4.39a), can be expressed as

$$[ds^2]_{2\mathrm{D},\chi}^{(k)} = \begin{cases} R^2(d\chi^2 + \sin^2\chi d\phi^2) & \text{for } k = +1, \\ R^2(d\chi^2 + \chi^2 d\phi^2) & \text{for } k = 0, \\ R^2(d\chi^2 + \sinh^2\chi d\phi^2) & \text{for } k = -1, \end{cases} \tag{4.40}$$

where we have factored out the overall scale R by introducing a dimensionless radial coordinate $\chi \equiv r/R$. Unlike the plane and sphere cases, there is no simple way to visualize this whole pseudosphere because the natural embedding is not into a flat space with Euclidean metric of $g_{ij} = \mathrm{diag}(1,1,1)$ but into a flat space with a pseudo-Euclidean metric of $g_{ij} = \mathrm{diag}(-1,1,1)$. Compared to the embedding of a sphere in a Euclidean space as (4.3), it can be worked out (see Problem 4.7) to show that the embedding of the $k = -1$ surface in such a pseudo-Euclidean space with coordinate (W, X, Y) corresponds to the condition

$$-W^2 + X^2 + Y^2 = -R^2. \tag{4.41}$$

While we cannot draw the whole pseudosphere in an ordinary 3D Euclidean space, the central portion of a saddle surface does represent a negative curvature surface, see Fig. 4.6(b).

In the cylindrical coordinates $(x^1, x^2) = (\rho, \phi)$, the metric for these three surfaces can be written in a particular compact form:

$$[ds^2]_{2\mathrm{D},\rho}^{(k)} = \frac{R^2 d\rho^2}{R^2 - k\rho^2} + \rho^2 d\phi^2. \tag{4.42}$$

We can easily check that for $k = 0$ the metric (4.42) yields $ds^2 = d\rho^2 + \rho^2 d\phi^2$, which is the infinitesimal separation for a flat surface with the familiar polar coordinates, cf. (4.33). For the positive curvature $k = +1$, the metric (4.42) is just the metric (4.15) of spherical surface in the cylindrical coordinate system.

3D spaces with constant curvature

The 2D spaces with constant curvature have metrics of (4.42) in the cylindrical coordinates and metrics of (4.40) in polar coordinates. We now make a heuristic argument for their generalization to 3D constant curvature spaces. Compared to the 2D coordinates (r, ϕ) or (ρ, ϕ), the 3D spherical coordinate system (r, θ, ϕ) or (ρ, θ, ϕ) involves an additional (polar) angle coordinate. Specifically for the $k = 0$ cases, we have the polar coordinate for a flat 2D surface, and the spherical coordinates for an Euclidean 3D space. Their respective metric relations are well-known:

$$[ds^2]^{(0)}_{\text{2D}} = dr^2 + r^2 d\phi^2 \tag{4.43}$$

and

$$[ds^2]^{(0)}_{\text{3D}} = dr^2 + r^2 d\Omega^2. \tag{4.44}$$

Namely, we replace the angular factor $d\phi^2$ by the solid angle factor $d\Omega^2 = (d\theta^2 + \sin^2\theta d\phi^2)$. Here we suggest that, even for the $k \neq 0$ spaces, we can obtain the 3D expressions in the same manner. From (4.40), we have the metric for 3D spaces in the spherical polar coordinates (χ, θ, ϕ)

$$[ds^2]^{(k)}_{\text{3D},\chi} = \begin{cases} R^2(d\chi^2 + \sin^2\chi d\Omega^2) & \text{for } k = +1, \\ R^2(d\chi^2 + \chi^2 d\Omega^2) & \text{for } k = 0, \\ R^2(d\chi^2 + \sinh^2\chi d\Omega^2) & \text{for } k = -1. \end{cases} \tag{4.45}$$

Similarly, if we replace the radial coordinate ρ by a dimensionless $\xi \equiv \rho/R$, we have, from (4.42), the metric for the (ξ, θ, ϕ) "cylindrical" coordinate system

$$[ds^2]^{(k)}_{\text{3D},\xi} = R^2\left(\frac{d\xi^2}{1 - k\xi^2} + \xi^2 d\Omega^2\right). \tag{4.46}$$

Eqs (4.45) and (4.46) reduce to the respective 2D metric expressions (4.40) and (4.42) when we take a 2D slice of the 3D space with either $d\theta = 0$ or $d\phi = 0$. This means that all the 2D subspaces are appropriately curved.

A rigorous derivation of these results would involve the mathematics of symmetric spaces, Killing vectors, and isometry. However, our heuristic deduction will be buttressed in Section 12.4.1 by a careful study of the properties of curvature tensor for a 3D space with the help of the Einstein equation.

The metrics in (4.45) and (4.46) with $k = +1$ describes a 3-sphere, $k = -1$ a 3-pseudosphere, and the overall distance scale R's are identified with the respective radii of these spheres. See Problem 4.6 for embedding of such 3D spaces in a 4D (pseudo-) Euclidean space—as generalizations of (4.3) and (4.41):

$$\pm W^2 + X^2 + Y^2 + Z^2 = \pm R^2 \tag{4.47}$$

with the plus sign for the space of 3-sphere, and negative sign the 3-pseudosphere. In Part II, we shall study cosmology based on the cosmological principle. The geometry of the cosmic spaces are the ones with constant curvature: $k = 0$ is the flat, $k = +1$ the closed, and $k = -1$ the open universes.

4.3.3 Curvature measures deviation from Euclidean relations

On a flat surface, the familiar Euclidean geometrical relations hold. For example, the circumference of a circle with radius r is $S = 2\pi r$, and the angular excess for any polygon equals to zero $\epsilon = 0$. The angular excess ϵ is defined to be the sum of the interior angles in excess of their flat space Euclidean value. For example, in the case of a triangle with angles α, β, and γ, the **angular excess** is defined as

$$\epsilon \equiv \alpha + \beta + \gamma - \pi. \tag{4.48}$$

The curvature measures how curved a surface is because it is directly proportional to the violation of Euclidean relations. In Fig. 4.6 we show two pictures of circles with radius r drawn on surfaces with nonvanishing curvature. It can be shown (Problem 4.9) that the circular circumference S differs from the flat surface value of $2\pi r$ by an amount controlled by the Gaussian curvature, K:

$$\lim_{r\to 0} \frac{2\pi r - S}{r^3} = \frac{\pi}{3} K. \tag{4.49}$$

For a positively curved surface the circumference is smaller than, for a negatively curved surface larger than, that on a flat space.

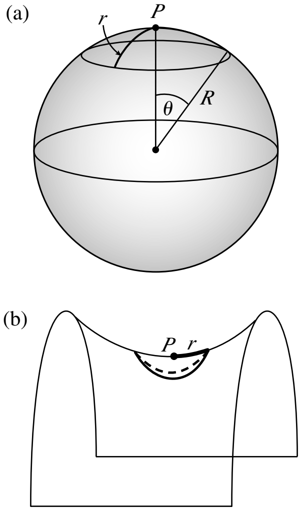

Fig. 4.6 A circle with radius r centered on point P, (a) on a spherical surface with curvature $K = 1/R^2$, (b) on the middle portion of a saddle shaped surface, which has negative curvature $K = -1/R^2$.

Angular excess and curvature

We shall also show that the angular excess ϵ is directly proportional to the area of the polygon σ with the proportional constant being the curvature K:

$$\epsilon = K\sigma. \tag{4.50}$$

This relation will be used in Chapter 11 to extract the general curvature, the Riemann curvature tensor, for a space of arbitrary dimensions. The contracted form of this Riemann tensor (called the Einstein tensor) enters directly in the GR field equation (the Einstein equation).

Here we shall explicitly prove (4.50) for the case of a spherical surface ($K = 1/R^2$). Let us first illustrate the validity of this relation for a particularly simple example of a triangle with two 90° interior angles and the third one being θ, as shown in Fig. 4.7(a). Clearly according to the definition of angular excess given in (4.48) we have $\epsilon = \theta$. The triangular area σ is exactly one-half of a **lune** with θ as its vertex angle. A lune is the area in between two great semi-circles, with an angle θ subtended between them, having an area value

$$\sigma_\theta = 2\theta R^2 \qquad \text{(area of a lune with angle } \theta\text{)}, \tag{4.51}$$

as $\theta = 2\pi$ corresponds to the whole spherical surface. Thus area of this triangle is $\sigma = \frac{1}{2}\sigma_\theta = \theta R^2$, which is just the relation (4.50) with $K = 1/R^2$.

The proof of (4.50) for a general triangle goes as follows. Draw three great circles (ABA′B′), (ACA′C′), and (BCB′C′) as in Fig. 4.7(b). Now consider the three lunes marked out by these geodesic lines, and record their respective areas according to (4.51):

$$\begin{aligned}
\sigma_\alpha &= 2\alpha R^2 && \text{(lune AA}' \text{ with angle } \alpha\text{)},\\
\sigma_\beta &= 2\beta R^2 && \text{(lune BB}' \text{ with angle } \beta\text{)},\\
\sigma_\gamma &= 2\gamma R^2 && \text{(lune CC}' \text{ with angle } \gamma\text{)}.
\end{aligned}$$

Their sum is

$$\sigma_\alpha + \sigma_\beta + \sigma_\gamma = 2(\alpha + \beta + \gamma)R^2. \tag{4.52}$$

However, an inspection of the diagram in Fig. 4.7(b) shows that the sum of these three lunes covers the entire front hemisphere, in addition to the triangular areas

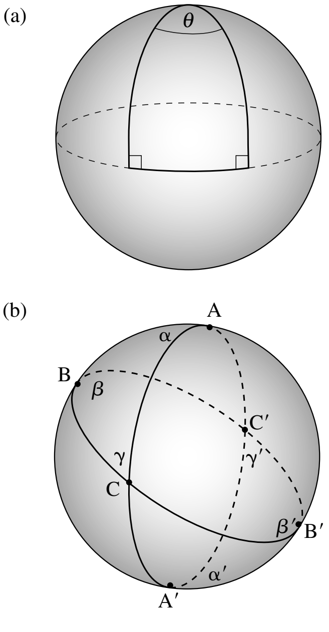

Fig. 4.7 (a) A triangle with two 90° interior angles on a spherical surface. (b) Three great circles ACA′C′, BCB′C′, and ABA′B′ intersect pairwise at points (A and A′), (B and B′), and (C and C′). The two identical triangles are ABC with angles α, β, γ on the front-hemisphere and A′B′C′ with angles α', β', γ' on the back-hemisphere.

of (ABC) and (A′B′C′). Thus another expression for the area sum is

$$\sigma_\alpha + \sigma_\beta + \sigma_\gamma = 2\pi R^2 + \sigma_{\mathrm{ABC}} + \sigma'_{\mathrm{ABC}}. \tag{4.53}$$

For spherical triangles, congruity of angles implies congruity of triangles themselves. Hence the angular equalities $\alpha = \alpha', \beta = \beta'$, and $\gamma = \gamma'$ imply the area equality of $\sigma_{\mathrm{ABC}} = \sigma'_{\mathrm{ABC}}$. Equations (4.52) and (4.53) then lead to

$$\alpha + \beta + \gamma = \pi + \sigma_{\mathrm{ABC}}/R^2. \tag{4.54}$$

Namely,

$$\alpha + \beta + \gamma - \pi = \epsilon = K\sigma_{\mathrm{ABC}}, \tag{4.55}$$

which is the claimed result (4.50) with $K = 1/R^2$.

Having demonstrated the validity of (4.50) for an arbitrary spherical triangle, it is not difficult to prove its validity for any spherical polygon (Problem 4.10). Furthermore, because a sufficiently small region on any curved 2D surface can be approximated by a spherical surface, (4.50) must hold for any infinitesimal polygon on any warped 2D space. In Section 11.3, this non-Euclidean relation will be used to generalize the notion of curvature (K) of a 2D space to that of an n-dimensional curved space.

From Gauss to Riemann

From Gauss's theory of curved surface, his student Bernhard Riemann showed that this algebraic approach to geometry can be extended to higher dimensional curved spaces when the spatial index ranges over $n = 1, 2, \ldots, n$ for an n-dimensional space. The virtue of this algebraic method is to make the study of higher dimensional non-Euclidean geometry, which by and large is impossible to visualize, more accessible. A few years before Riemann's presentation of his result in 1854, Bólyai and Lobachevsky had already introduced non-Euclidean geometry—the geometry without the parallelism axiom. However, their work remained unappreciated until Riemann showed that his larger framework encompassed the Bólyai and Lobachevsky results. This just shows the power of the Riemannian approach.

Our interest will mostly be the 4D spacetime with the index $\mu = 0, 1, 2, 3$. For such a 1 + 3-dimensional manifold, the flat metric corresponds to $g_{\mu\nu} = \eta_{\mu\nu} = \mathrm{diag}(-1, 1, 1, 1)$. We should also mention that the extension to higher dimensional space is nontrivial, because, beyond two dimensions, the curvature of the space can no longer be described by a single function.

Review questions

1. What does it mean to have an "intrinsic geometric description" (vs. "extrinsic description")?
2. Provide a description of the intrinsic geometric operations to fix the metric elements.
3. How does the geodesic equation represent the curve $x^\alpha(\sigma)$ having an extremum length? (Just say it in words the relation of the geodesic equation to the extremum length condition.)
4. In what sense does the metric function describe all the intrinsic geometric properties of a space? Namely, is the metric an intrinsic quantity? What is the relation between other intrinsic geometric quantities and the metric?

5. Curved surfaces necessarily have a position-dependent metric. But it is not a sufficient condition. Illustrate this point with an example.

6. What is the fundamental difference between the coordinate transformations in a curved space and those in flat space (e.g. Lorentz transformations in the flat Minkowski space)?

7. What is the "flatness theorem"?

8. In what sense is the Gaussian curvature K a good criterion for finding out whether a surface is curved or not?

9. What are the three surfaces of constant curvature?

10. Write out its embedding equation in a 4D space of a 3-sphere as well as the equation of a 3-pseudosphere. Is the embedding space for the latter an Euclidean space?

11. The curvature measures how curved a space is because it controls the amount of deviation from Euclidean relations. Give an example of such a non-Euclidean relation, showing the deviation from flatness being proportional to the curvature.

12. What is angular excess? How is it related to the Gaussian curvature? Give a simple example of a polygon on a spherical surface that clearly illustrates this relation.

Problems

(4.1) **Metric for the spherical surface in cylindrical coordinates** Show that the metric for the spherical surface with "cylindrical coordinates" of (4.15) follows from (4.6) and the Pythagorean relation $ds^2 = dX^2 + dY^2 + dZ^2$ in the embedding space.

(4.2) **Basis vectors on a spherical surface** Equations (4.7) and (4.9) are, respectively, two equivalent and closely related definitions of the metric. In the text we use (4.9) to deduce the metric matrix (4.13) for a spherical surface. What are the corresponding basis vectors for this coordinate system? Check that they yield the same matrix through the definition of (4.7).

(4.3) **Coordinate transformation of the metric** Use (2.45) to show explicitly that the transformation (4.18) relates the metric tensors of the two coordinate systems as given in (4.13) and (4.15).

(4.4) **Geodesics on simple surfaces** Use the geodesic Eq. (4.30) to confirm the familiar results that the geodesic is (a) a straight line on a flat plane and (b) a great circle on a spherical surface.

(4.5) **Locally flat metric** Explicitly display a transformation that turns a general 2D metric **at a point** to the Euclidean metric $\bar{g}_{ab} = \delta_{ab}$, the Kronecker delta, or the pseudo-Euclidean metric $\bar{g}_{ab} = \eta_{ab}$, where $\eta_{ab} = \text{diag}(1, -1)$. (There are an infinite number of such transformations, just display one.)

(4.6) **Checking the Gaussian curvature formula** Checking the connection of the metric and the curvature for the three surfaces with constant curvature $K = k/R^2$ by plugging (4.42) and (4.40) into the expression for the curvature in (4.35).

(4.7) **3-sphere and 3-pseudosphere**

(a) **3D flat space** Express Cartesian coordinates (x, y, z) in terms of polar coordinates (r, θ, ϕ). Show that the solid angle factor in polar coordinate expression for the infinitesimal separation satisfies the relation $r^2 d\Omega^2 = dx^2 + dy^2 + dz^2 - dr^2$.

(b) **3-sphere** Consider the possibility of embedding a 3-sphere in a 4D Euclidean space with Cartesian coordinates (W, X, Y, Z). From (4.45) for 3-sphere $(k = +1)$ and an expression relating Cartesian coordinates to solid angle differential $d\Omega^2$ as suggested by the equation shown in part (a) to find the differential for the new coordinate dW. This result should suggest that the 4D embedding space is indeed Euclidean with $W = R\cos(r/R)$. From this, display the entire expression for (W, X, Y, Z) in terms of polar coordinates (r, θ, ϕ). Furthermore, verify that the 3-sphere is a 3D subspace satisfying the constraint

$$W^2 + X^2 + Y^2 + Z^2 = R^2.$$

(c) **3-pseudosphere** Now the fourth coordinate should be $W = R\cosh(r/R)$. Show that the 4D embedding space is pseudo-Euclidean with a metric $\eta_{\mu\nu} = \text{diag}(-1, 1, 1, 1)$ and the 3-pseudosphere is a 3D subspace satisfying the condition

$$-W^2 + X^2 + Y^2 + Z^2 = -R^2.$$

(4.8) **Volume of higher dimensional space** The general expression for the differential volume is the product of coordinate differentials **and** the square root of the metric determinant:

$$dV = \sqrt{\det g} \prod_i dx^i. \tag{4.56}$$

(a) Verify that for 3D flat space this reduces to the familiar expression for volume element $dV = dx\,dy\,dz$ in the Cartesian coordinates, and $dV = r^2 \sin\theta dr d\theta d\phi$ in the spherical coordinates.

(b) Work out the volume element for a 3-sphere, and integrate it to obtain the result of $V_3^{(+1)} = 2\pi^2 R^3$. Thus 3-sphere, much like the familiar spherical surface, is a 3D space having no boundary yet with a finite volume. When applied to cosmology this is a "closed universe" with R being referred to as "the radius of the universe."

(4.9) **Non-Euclidean relation between radius and circumference of a circle** On a curved surface the circumference S of a circle is no longer related to its radius r by $S = 2\pi r$. The deviation from this flat space relation is proportional to the curvature, as shown in (4.49). Derive this relation for the simple cases (a) a sphere and (b) a pseudosphere.

(4.10) **Angular excess and polygon area** Generalize the proof of (4.50) to the case of an arbitrary polygon. Namely, one still has $\epsilon = \sigma K$ with ϵ being the angular excess over the Euclidean sum of the polygon and σ being the area of the polygon.

GR as a geometric theory of gravity - I

5

- We first present a *geometric* description of equivalence principle (EP) physics of gravitational time dilation. In this geometric theory, the metric $g_{\mu\nu}(x)$ plays the role of relativistic gravitational potential.
- Curved spacetime being the gravitational field, geodesic equation in spacetime is the GR equation of motion, which is checked to have the correct Newtonian limit.
- At every spacetime point, one can construct a free-fall frame in which gravity is transformed away. However, in a finite-sized region, one can detect the residual tidal force which are second derivatives of gravitational potential. It is the curvature of spacetime.
- The GR field equation directly relates the mass/energy distribution to spacetime's curvature. Its solution is the metric function $g_{\mu\nu}(x)$, determining the geometry of spacetime.

In Chapter 3 we have deduced several pieces of physics from the empirical principle of equivalence of gravity and inertia. In Chapter 4, elements of the mathematical description of a curved space have been presented. In this chapter, we shall show how some of equivalence principle (EP) physics can be interpreted as the geometric effects of curved spacetime. Such study motivated Einstein to propose his general theory of relativity, which is a geometric theory of gravitation, with equation of motion being the geodesic equation, and field equation in the form of the curvature being proportional to the mass/energy source fields.

5.1 Geometry as gravity

By a geometric theory, or a geometric description, of any physical phenomenon we mean that the physical measurement results can be attributed directly to the underlying geometry of space and time. This is illustrated by the example we discussed in Section 4.2 in connection with a spherical surface as shown in Fig. 4.2. The length measurements on the surface of earth are different in different directions: the east and west distances between any pairs of points separated by the same azimuthal angle $\Delta\phi$ become smaller as they move away from the equator, while the lengths in the north and south directions for a fixed ϕ remain the same, we could, in principle, interpret such results in two equivalent ways:

1. Without considering that the 2D space is curved, we can say that physics (i.e. dynamics) is such that the measuring ruler changed scale when

pointing in different directions—much in the same manner FitzGerald–Lorentz length contraction was originally interpreted by physicists in this manner.

2. The alternative description (the "geometric theory") is that we use a standard ruler with a fixed scale (defining the coordinate distance) and the varying length measurements are attributed to the underlying geometry of a curved spherical surface. This is expressed mathematically in the form of a position-dependent metric tensor $g_{ab}(x) \neq \delta_{ab}$.

Einstein's general theory of relativity is a geometric theory of gravity—gravitational phenomena are attributed as reflecting the underlying curved spacetime. An invariant (with respect to coordinate transformations) interval is related to coordinates of the spacetime manifold through the metric in the form of

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu. \tag{5.1}$$

The Greek indices range over $(0, 1, 2, 3)$ with $x^0 = ct$ and the metric $g_{\mu\nu}$ is a 4×4 matrix. Observers measure with rulers and clocks. Thus the spacetime manifold not only expresses the spatial relations among events but also their causal structure. For special relativity (SR) we have the geometry of a flat spacetime with a position-independent metric $g_{\mu\nu} = \eta_{\mu\nu} = \text{diag}(-1, 1, 1, 1)$. GR as a geometric theory of gravity posits that matter and energy cause spacetime to warp $g_{\mu\nu} \neq \eta_{\mu\nu}$, and gravitational phenomena are just the effects of a curved spacetime on a test object.

How did the study of the physics as implied by the equivalence principle (EP) motivate Einstein to propose that the relativistic gravitational field was the curved spacetime? We have already discussed the EP physics of gravitational time dilation—clocks run at different rates at positions having different gravitational potential values $\Phi(\mathbf{x})$, as summarized in (3.32). This variation of time rate follows a definite pattern. Instead of working with a complicated scheme of clocks running at different rates, this physical phenomenon can be given a geometric interpretation as showing a nontrivial metric, $g_{\mu\nu} \neq \eta_{\mu\nu}$. Namely, a simpler way of describing the same physical situation is by using a stationary clock at $\Phi = 0$ as the standard clock. Its fixed rate is taken to be the time coordinate t. One can then compare the time intervals $d\tau(\mathbf{x})$ measured by clocks located at other locations (the proper time interval at $\mathbf{x}$) to this coordinate interval dt. According to EP as stated in (3.38), we should find

$$d\tau(\mathbf{x}) = \left(1 + \frac{\Phi(\mathbf{x})}{c^2}\right) dt. \tag{5.2}$$

The geometric approach says that the measurement results can be interpreted as showing a spacetime with a warped geometry having a metric element of

$$g_{00} = -\left(1 + \frac{\Phi(x)}{c^2}\right)^2 \simeq -\left(1 + \frac{2\Phi(x)}{c^2}\right). \tag{5.3}$$

This comes about because (5.1) reduces down to $ds^2 = g_{00} dx^0 dx^0$ for $d\mathbf{x} = 0$, as appropriate for proper time interval (rest frame time interval) and the knowledge that the invariant interval is just the proper time interval $ds^2 = -c^2 d\tau^2$, leading to the expression

$$(d\tau)^2 = -g_{00}(dt)^2. \tag{5.4}$$

The result in (5.3) states that the metric element g_{00} in the presence of gravity deviates from the flat spacetime value of $\eta_{00} = -1$ because of the presence of gravity. Thus the geometric interpretation of the EP physics of gravitational time dilation is to say that gravity changes the spacetime metric element g_{00} from -1 to an x-dependent function. Gravity warps spacetime—in this case warps it in the time direction. Also, since g_{00} is directly related to the Newtonian gravitational potential $\Phi(x)$ as in (5.3), we can say that the ten independent components of the spacetime metric $g_{\mu\nu}(x)$ are the "relativistic gravitational potentials."

5.1.1 EP physics and a warped spacetime

Adopting a geometric interpretation of EP physics, we find that resultant geometry has all the characteristic features of a warped manifold of space and time: a position-dependent metric, deviations from Euclidean geometric relations, and at every location we can always transform gravity away to obtain a flat spacetime, just as one can always find a locally flat region in a curved space.

Position-dependent metrics

As we have discussed in Section 4.2, the metric tensor in a curved space is necessarily position-dependent. Clearly, (5.3) has this property. In Einstein's geometric theory of gravitation, the metric function is all that we need to describe the gravitational field completely. $g_{\mu\nu}(x)$ plays the role of relativistic gravitational potentials, just as $\Phi(x)$ is the Newtonian gravitational potential.

Non-Euclidean relations

In a curved space Euclidean relations no longer hold (cf. Section 4.3.3), for example, the sum of interior angles of a triangle on the surface of spherical surface deviates from π, the ratio of circular circumference to the radius is different from the value of 2π. As it turns out, EP does imply non-Euclidean relation among geometric measurements. We illustrate this with a simple example. Consider a cylindrical room in high speed rotation around its axis. This acceleration case, according to EP, is equivalent to a centrifugal gravitational field. (This is one way to produce "artificial gravity.") For such a rotating frame, one finds that, because of special relativistic (longitudinal) length contraction, 2π times the radius, which is not changed because velocity is perpendicular to the radial direction, will no longer equal the circular circumference of the cylinder, cf. Fig. 5.1 and Problem 5.3. Thus Euclidean geometry is no longer valid in the presence of gravity. We reiterate this connection: the rotating frame, according to EP, is a frame with gravity; the rotating frame, according to SR length contraction, has a relation between its radius and circumference that is not Euclidean. Hence, we say the presence of gravity brings about non-Euclidean geometry. (Distance measurement in a curved spacetime is discussed in Problem 5.2.)

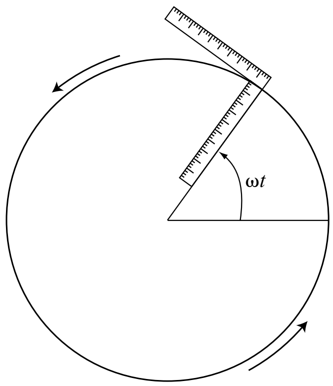

Fig. 5.1 Rotating cylinder with length contraction in the tangential direction but not in the radial direction, resulting in a non-Euclidean relation between circumference and radius.

Local flat metric and local inertial frame

In a curved space a small local region can always be described approximately as a flat space. A more precise statement is given by the flatness theorem of Section 4.2.2. Now, if we identify our spacetime as the gravitational field, is the corresponding flatness theorem valid? The answer is in the affirmative. Actually this is the essence of EP, stating that we can always transform gravity away in a local region. In this region, because of the absence of gravity, SR is valid and the metric is the flat Minkowski metric. General relativity (GR) has the same local lightcone structure as SR: $ds^2 < 0$ being timelike, $ds^2 > 0$ spacelike, and $ds^2 = 0$ lightlike. The relation between local flat and local inertial frames will be further explored in Section 5.3, where we show that the spacetime curvature is the familiar tidal force.

5.1.2 Curved spacetime as gravitational field

Recall that a field theoretical description of the interaction between a source and a test particle is a two-step description:

$$\boxed{\text{Source particle}} \quad \xrightarrow[\text{equation}]{\text{Field}} \quad \boxed{\text{Field}} \quad \xrightarrow[\text{motion}]{\text{Equation of}} \quad \boxed{\text{Test particle}}$$

Instead of the source particle acting directly on the test particle through some instantaneous action-at-a-distance force, the source creates a field everywhere, and the field then acts on the test particle locally. The first-step is given by the field equation which, given the source distribution, determines the field everywhere. In the case of electromagnetism it is Maxwell's equation. The second-step is provided by the equation of motion, which allows us to find the motion of the test particle, once the field function is known. The electromagnetic equation of motion follows directly from the Lorentz force law.

Newtonian gravitational field

The field equation in Newton's theory of gravity, when written in terms of the gravitational potential $\Phi(x)$, is given by (3.6)

$$\nabla^2\Phi = 4\pi G_{\rm N}\rho, \tag{5.5}$$

where $G_{\rm N}$ is Newton's constant, and ρ is the mass density function. The Newtonian theory is not a dynamic field theory as it does not provide a description of time evolution. Namely, it is the static limit of some field theory, thus has no field propagation. The Newtonian equation of motion is Eq. (3.8)

$$\frac{d^2\mathbf{r}}{dt^2} = -\boldsymbol{\nabla}\Phi. \tag{5.6}$$

The task Einstein undertook was to find the relativistic generalizations of these two sets of Eqs (5.5) and (5.6). Since in relativity, space and time are treated on equal footing, a successful relativistic program will automatically yield a dynamical theory as well.

Relativistic gravitational field

The above discussion suggests that the EP physics can be described in geometric language. The resultant mathematics coincides with that describing a warped spacetime. Thus it is simpler, and more correct, to say that relativistic gravitational field **is** the curved spacetime. The effect of the gravitational interaction between two particles can be described as the source mass giving rise to a curved spacetime which in turn influences the motion of the test mass. Or, put more strongly, EP requires a metric structure of spacetime and particles follow geodesics in such a curved spacetime.

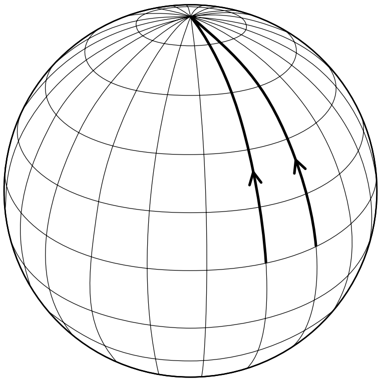

Fig. 5.2 Two particle trajectories with decreasing separation can be interpreted either as resulting from an attractive force or as reflecting the underlying geometry of a spherical surface.

The possibility of using a curved space to represent a gravitational field can be illustrated with the following example involving a 2D curved surface. Two masses on a spherical surface start out at the equator and move along two geodesic lines as represented by the longitudinal great circles. As they move along, the distance between them decreases (Fig. 5.2). We can attribute this to some attractive force between them, or simply to the curved space causing their trajectory to converge. That is to say, this phenomenon of two convergent particle trajectories can be thought of either as resulting from an attractive tidal force, or from the curvature of the space.[1] Eventually we shall write down the relativistic gravitational equations. In Einstein's approach these differential equations can be thought of as reflecting an underlying warped spacetime.

Based on the study of EP phenomenology, Einstein made the conceptual leap (a logical deduction, but a startling leap nevertheless) to the idea that curved spacetime **is** the gravitational field:

$$\boxed{\text{Source}} \xrightarrow[\substack{\text{Einstein field}\\ \text{equation}}]{} \boxed{\text{Curved spacetime}} \xrightarrow[\substack{\text{Geodesic}\\ \text{equation}}]{} \boxed{\text{Test particle}}$$

The mass/energy source gives rise to a warped spacetime, which in turn dictates the motion of the test particle. Plausibly the test particle moves along the shortest and straightest possible curve in the curved manifold. Such a line is the geodesic curve. Hence the GR equation of motion is the geodesic equation (Section 5.2). The GR field equation is the Einstein equation, which relates the mass/energy distribution to the curvature of spacetime (Section 5.3).

In this way GR fulfills Einstein's conviction that "space is not a thing": the ever changing relation of matter and energy is reflected by an ever changing geometry. Spacetime does not have an independent existence; it is nothing but an expression of the relations among physical processes in the world.

5.2 Geodesic equation as GR equation of motion

The metric function $g_{\mu\nu}(x)$ in (5.1) describes the geometry of curved spacetime. In GR the mass/energy source determines the metric function through the field equation. Namely, $g_{\mu\nu}(x)$ is the solution of the GR field equation. Knowing $g_{\mu\nu}(x)$ one can write down the equation of motion, which fixes the trajectory of the test particle. It is natural to expect the test particle to follow in this spacetime the shortest and straightest possible trajectory, the **geodesic curve**. Thus, GR equation of motion should coincide with the geodesic equation. In Section 4.2.1, we have derived the geodesic equation from the property of the geodesic line as the curve with extremum length. We also recall that a point

[1] Further reference to gravitational tidal forces vs. curvature description of the relative separation between two particle trajectories can be found in Section 5.3.1 when we discuss the Newtonian deviation equation for tidal forces, which has its GR generalization "equation of geodesic deviation," given in Chapter 12, cf. Problems 12.4 and 12.5.

in spacetime is an event and that the trajectory is a worldline (cf. Box 5.1). The geodesic equation determines the worldline that a test particle will follow under the influence of gravity. The geodesic equation in spacetime is Eq. (4.30) with its Latin indices $a = 1, 2$ changed into Greek indices $\mu = 0, 1, 2, 3$ with $x^0 = ct$.

$$\frac{d}{d\sigma} g_{\mu\nu}\dot{x}^{\nu} - \frac{1}{2}\frac{\partial g_{\lambda\rho}}{\partial x^{\mu}}\dot{x}^{\lambda}\dot{x}^{\rho} = 0, \tag{5.7}$$

where $x^{\mu} = x^{\mu}(\sigma)$ with σ being the curve parameter, and $\dot{x}^{\mu} \equiv dx^{\mu}/d\sigma$.

We can cast (5.7) into a more symmetric form which will also facilitate our later interpretation (in Section 11.2.2) of the geodesic as the straightest possible curve. Carrying out the differentiation of the first term and noting that the metric's dependence on σ is entirely through $x^{\mu}(\sigma)$:

$$g_{\mu\nu}\frac{d^2x^{\nu}}{d\sigma^2} + \frac{\partial g_{\mu\nu}}{\partial x^{\lambda}}\frac{dx^{\lambda}}{d\sigma}\frac{dx^{\nu}}{d\sigma} - \frac{1}{2}\frac{\partial g_{\lambda\rho}}{\partial x^{\mu}}\frac{dx^{\lambda}}{d\sigma}\frac{dx^{\rho}}{d\sigma} = 0. \tag{5.8}$$

Since the product $(dx^{\lambda}/d\sigma)(dx^{\nu}/d\sigma)$ in the second term is symmetric with respect to the interchange of indices λ and ν, only the symmetric part of its coefficient:

$$\frac{1}{2}\left(\frac{\partial g_{\mu\nu}}{\partial x^{\lambda}} + \frac{\partial g_{\mu\lambda}}{\partial x^{\nu}}\right)$$

can enter. In this way the geodesic Eq. (5.7) can be cast (after relabeling some repeated indices) into the form,

$$\frac{d^2x^{\nu}}{d\sigma^2} + \Gamma^{\nu}_{\lambda\rho}\frac{dx^{\lambda}}{d\sigma}\frac{dx^{\rho}}{d\sigma} = 0, \tag{5.9}$$

where

$$g_{\mu\nu}\Gamma^{\nu}_{\lambda\rho} = \frac{1}{2}\left[\frac{\partial g_{\lambda\mu}}{\partial x^{\rho}} + \frac{\partial g_{\rho\mu}}{\partial x^{\lambda}} - \frac{\partial g_{\lambda\rho}}{\partial x^{\mu}}\right]. \tag{5.10}$$

$\Gamma^{\nu}_{\lambda\rho}$, being this particular combination of the first derivatives of the metric tensor, is called the **Christoffel symbol** (also known as the **connection**). The geometric significance of this quantity will be studied in Chapter 11. From now on we will use the geodesic equation in the form as given in (5.9). To reiterate, the geodesic equation is the equation of motion in GR because it is the shortest curve in a warped spacetime. By this we mean that once the gravitation field is given, that is, spacetime functions $g_{\mu\nu}(x)$ and $\Gamma^{\mu}_{\nu\lambda}(x)$ are known, (5.9) tells us how a test particle will move in such a field: it will always follow the shortest and the straightest possible trajectory in this spacetime. A fuller justification of using the geodesic equation as the GR equation of motion will be given in Section 12.1.1.

5.2.1 The Newtonian limit

Here we shall show that the geodesic Eq. (5.9) as the GR equation of motion, reduces to the Newtonian equation of motion (5.6) in the "Newtonian limit" of a test particle moving with nonrelativistic velocity $v \ll c$ in a static and weak gravitational field.

Nonrelativistic speed $(dx^i/dt) \ll c$. This inequality $dx^i \ll cdt$ implies that

$$\frac{dx^i}{d\sigma} \ll c\frac{dt}{d\sigma} = \frac{dx^0}{d\sigma}. \tag{5.11}$$

Keeping only the dominant term $(dx^0/d\sigma)(dx^0/d\sigma)$ in the double sum over indices λ and ρ of the geodesic Eq. (5.9), we have

$$\frac{d^2x^\mu}{d\sigma^2} + \Gamma^\mu_{00}\frac{dx^0}{d\sigma}\frac{dx^0}{d\sigma} = 0. \tag{5.12}$$

Static field $(\partial g_{\mu\nu}/\partial x^0) = 0$. Because all time derivatives vanish, the Christoffel symbol of (5.10) takes a simpler form

$$g_{\nu\mu}\Gamma^\mu_{00} = -\frac{1}{2}\frac{\partial g_{00}}{\partial x^\nu}. \tag{5.13}$$

Weak field $h_{\mu\nu} \ll 1$. We assume that the metric is not too different from the flat spacetime metric $\eta_{\mu\nu} = \text{diag}(-1, 1, 1, 1)$

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}, \tag{5.14}$$

where $h_{\mu\nu}(x)$ is a small correction field. Keeping in mind that flat space has a constant metric $\eta_{\mu\nu}$, we have $\partial g_{\mu\nu}/\partial x^\lambda = \partial h_{\mu\nu}/\partial x^\lambda$ and the Christoffel symbol is of order $h_{\mu\nu}$. To the leading order, (5.13) is

$$\eta_{\nu\mu}\Gamma^\mu_{00} = -\frac{1}{2}\frac{\partial h_{00}}{\partial x^\nu}, \tag{5.15}$$

which because $\eta_{\nu\mu}$ is diagonal has, for a static h_{00}, the following components

$$-\Gamma^0_{00} = -\frac{1}{2}\frac{\partial h_{00}}{\partial x^0} = 0 \quad \text{and} \quad \Gamma^i_{00} = -\frac{1}{2}\frac{\partial h_{00}}{\partial x^i}. \tag{5.16}$$

We can now evaluate (5.12) by using (5.16): the $\mu = 0$ part leads to

$$\frac{dx^0}{d\sigma} = \text{constant} \tag{5.17}$$

and the $\mu = i$ part is

$$\frac{d^2x^i}{d\sigma^2} + \Gamma^i_{00}\frac{dx^0}{d\sigma}\frac{dx^0}{d\sigma} = \left(\frac{d^2x^i}{c^2dt^2} + \Gamma^i_{00}\right)\left(\frac{dx^0}{d\sigma}\right)^2 = 0, \tag{5.18}$$

where we have used (5.17) to go from $(d^2x^i/d\sigma^2)$ to $(d^2x^i/dx^{0\,2})(dx^0/d\sigma)^2$. The above equation, together with (5.16), implies

$$\frac{d^2x^i}{c^2dt^2} - \frac{1}{2}\frac{\partial h_{00}}{\partial x^i} = 0, \tag{5.19}$$

which is to be compared with the Newtonian equation of motion (5.6). Thus $h_{00} = -2\Phi/c^2$ and using the definition of (5.14) we recover (5.3) first obtained heuristically in Section 5.1:

$$g_{00} = -\left(1 + \frac{2\Phi(x)}{c^2}\right). \tag{5.20}$$

We can indeed regard the metric tensor as the relativistic generalization of the gravitational potential. This expression also provides us with a criterion to characterize a field being weak as in (5.14):

$$\left[|h_{00}| \ll |\eta_{00}|\right] \Rightarrow \left[\frac{|\Phi|}{c^2} \ll 1\right]. \tag{5.21}$$

Consider the gravitational potential at earth's surface. It is equal to the gravitational acceleration times earth's radius, $\Phi_\oplus = g \times R_\oplus = O\ (10^7 \mathrm{m}^2/\mathrm{s}^2)$, or $\Phi_\oplus/c^2 = O\ (10^{-10})$. Thus a weak field is any gravitational field being less than 10 billion $g's$.

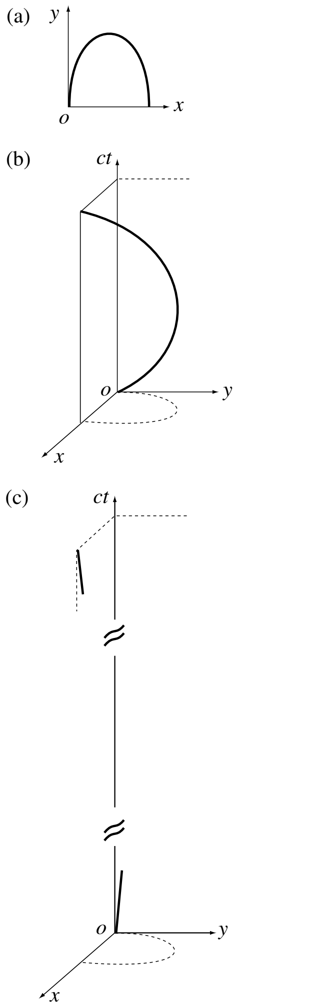

Fig. 5.3 (a) Particle trajectory in the (x, y) plane. (b) Particle worldline with projection onto the (x, y) plane as shown in (a). (c) Spacetime diagram with the time axis stretched a great distance.

Box 5.1 The geodesic is the worldline of a test-particle

It may appear somewhat surprising to hear that a test particle will follow a "straight line" in the presence of a gravitational field. After all, our experience is just the opposite: when we throw an object, it follows a parabolic trajectory. Was Einstein saying that the parabolic trajectory is actually straight? All such paradoxes result from confusing the 4D spacetime with the ordinary 3D space. The GR equation of motion tells us that a test particle will follow a geodesic line in spacetime, which is not a geodesic line in the 3D space. Namely, the worldline of a particle should be a geodesic, which generally does not imply a straight trajectory in the spatial subspace. A simple illustration using the space–time diagram should make this clear.

Let us consider the case of throwing an object to a height of 10 m over a distance of 10 m. Its spatial trajectory is displayed in Fig. 5.3(a). When we represent the corresponding worldline in the spacetime diagram we must plot the time-axis *ct* also, see Fig. 5.3(b). For the case under consideration, this object takes 1.4 s to reach the highest point and another 1.4 s to come down. But a 2.8-s interval will be represented by almost 1 million kilometers of *ct* in the spacetime diagram (more than the round trip distance to the moon). When the time axis is stretched out in this way, one then realizes this worldline is very straight indeed, see Fig. 5.3(c). The straightness of this worldline reflects the fact that the terrestrial gravity is a very weak field (recall $\Phi_\oplus/c^2 \simeq 10^{-10}$)—it curves the spacetime only a tiny amount. Namely, in this case the spacetime is practically flat, and thus the geodesic is very close to a straight line.

5.2.2 Gravitational redshift revisited

Previously in Chapter 3 we have shown that the strong EP implied a gravitational redshift (in a static gravitational field) of frequency ω

$$\frac{\Delta\omega}{\omega} = -\frac{\Delta\Phi}{c^2}. \tag{5.22}$$

From this result we heuristically deduced that, in the presence of nonzero gravitational potential, the metric must deviate from the flat space value. Now in this chapter, we have seen that Einstein's theory based on a curved spacetime has the result (5.20) in the Newtonian limit. This, as shown in (5.2), can be stated as a relation between the proper time τ and the coordinate time t

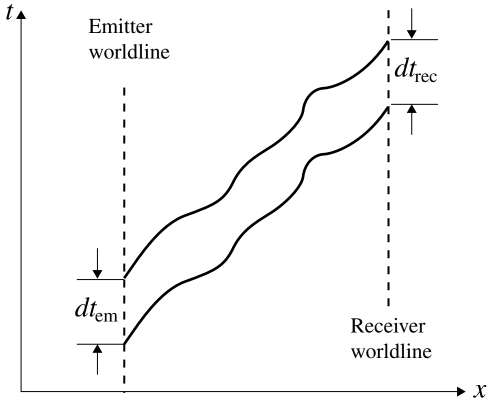

Fig. 5.4 Worldlines for two light wavefronts propagating from emitter to receiver in a static curved spacetime.

as follows:

$$d\tau = \sqrt{-g_{00}}\,dt \qquad \text{with } g_{00} = -\left(1 + 2\frac{\Phi}{c^2}\right). \tag{5.23}$$

Here we wish to see how the gravitational frequency shift result of (5.22) emerges in this curved spacetime description.

In Fig. 5.4, the two curved lines are the lightlike worldlines of two wavefronts emitted at an interval $dt_{\rm em}$ apart. They are curved because in presence of gravity the spacetime is curved. (In the flat spacetime, they would be two straight 45°-lines.) Because we are working with static gravitational field (hence a time-independent spacetime curvature), this $dt_{\rm em}$ time interval between the two wavefronts is maintained throughout the trip until they are received. Namely, these two wavefronts trace out two congruent worldlines. In particular, the coordinate time separations at emission and reception are identical,

$$dt_{\rm em} = dt_{\rm rec}. \tag{5.24}$$

On the other hand, the frequency being inversely proportional to the proper time interval $\omega = 1/d\tau$, we can then use (5.23) and (5.24) to derive:

$$\begin{aligned}\frac{\omega_{\rm rec}}{\omega_{\rm em}} &= \frac{d\tau_{\rm em}}{d\tau_{\rm rec}} = \frac{\sqrt{-(g_{00})_{\rm em}}\,dt_{\rm em}}{\sqrt{-(g_{00})_{\rm rec}}\,dt_{\rm rec}} = \left(\frac{1 + 2(\Phi_{\rm em}/c^2)}{1 + 2(\Phi_{\rm rec}/c^2)}\right)^{1/2} \\ &= 1 + \frac{\Phi_{\rm em} - \Phi_{\rm rec}}{c^2} + O\left(\frac{\Phi^2}{c^4}\right),\end{aligned} \tag{5.25}$$

which is the claimed result of (5.22):

$$\frac{\omega_{\rm rec} - \omega_{\rm em}}{\omega_{\rm em}} = \frac{\Phi_{\rm em} - \Phi_{\rm rec}}{c^2}. \tag{5.26}$$

5.3 The curvature of spacetime

We have already discussed in Chapter 4 (see especially Section 4.2.3) that in a curved space each small region can be approximated by a flat space, that is, locally a metric can always be approximated by a flat space metric.

This coordinate-dependence of the metric shows that the metric value cannot represent the essential feature of a curved space. However, as shown in Section 4.3 (and further discussion in Section 11.3), there exits a mathematical quantity involving the second derivative of the metric, called the curvature, which does represent the essence of a curved space: the space is curved if and only if the curvature is nonzero; and, also, the deviations from Euclidean relations are always proportional to the curvature.

If the warped spacetime is the gravitational field, what then is its curvature? What is the physical manifestation of this curvature? How does it enter in the GR equations of gravitation?

5.3.1 Tidal force as the curvature of spacetime

The equivalence principle states that in a freely falling reference frame the physics is the same as that in an inertial frame with no gravity. SR applies and the metric is given by the Minkowski metric $\eta_{\mu\nu}$. As shown in the flatness theorem (Section 4.2.2), this approximation of $g_{\mu\nu}$ by $\eta_{\mu\nu}$ can be done **only** locally, that is, in an appropriately small region. Gravitational effects can always be detected in a **finite-sized** free-fall frame as gravitational field is never strictly uniform in reality; the second derivatives of the metric come into play.

Consider the lunar gravitation attraction exerted on the earth. While the earth is in free fall toward the moon and *viceversa*, there is still a detectable lunar gravitational effect on earth. It is so because different points on earth will feel slightly different gravitational pulls by the moon, as depicted in Fig. 5.5(a). The center-of-mass (CM) force causes the earth to “fall towards the moon” so that this CM gravitational effect is “cancelled out” in this freely falling terrestrial frame. After subtracting out this CM force, the remanent forces on the earth, as shown in Fig. 5.5(b), are stretching in the longitudinal direction and compression in the transverse direction. They are just the familiar tidal forces.[2] Namely, in the freely falling frame, the CM gravitational effect is transformed away, but, there are still the remnant tidal forces. They reflect the **differences** of the gravitational effects on neighboring points, and are thus proportional to the derivative of the gravitational field. We can illustrate this point by the following observation. With r_{s} and r_{m} being the distances from earth to the sun and moon, respectively, we have

$$\left[\mathfrak{g}_{\mathrm{s}} = \frac{G_{\mathrm{N}}M_{\odot}}{r_{\mathrm{s}}^2}\right] > \left[\mathfrak{g}_{\mathrm{m}} = \frac{G_{\mathrm{N}}M_{\mathrm{m}}}{r_{\mathrm{m}}^2}\right], \tag{5.27}$$

showing that the gravitational attraction of the earth by the sun is much larger than that by the moon. On the other hand, because the tidal force is given by the derivative of the force field

$$\left[\frac{\partial}{\partial r}\frac{G_{\mathrm{N}}M}{r^2}\right] \propto \left[\frac{G_{\mathrm{N}}M}{r^3}\right], \tag{5.28}$$

and because $r_{\mathrm{s}} \gg r_{\mathrm{m}}$, the lunar tidal forces nevertheless end up being stronger than the solar ones

$$\left[\mathfrak{T}_{\mathrm{s}} = \frac{G_{\mathrm{N}}M_{\odot}}{r_{\mathrm{s}}^3}\right] < \left[\mathfrak{T}_{\mathrm{m}} = \frac{G_{\mathrm{N}}M_{\mathrm{m}}}{r_{\mathrm{m}}^3}\right]. \tag{5.29}$$

[2] The ocean is pulled away in opposite directions giving rise to two tidal bulges. This explains why, as the earth rotates, there are two high tides in a day.

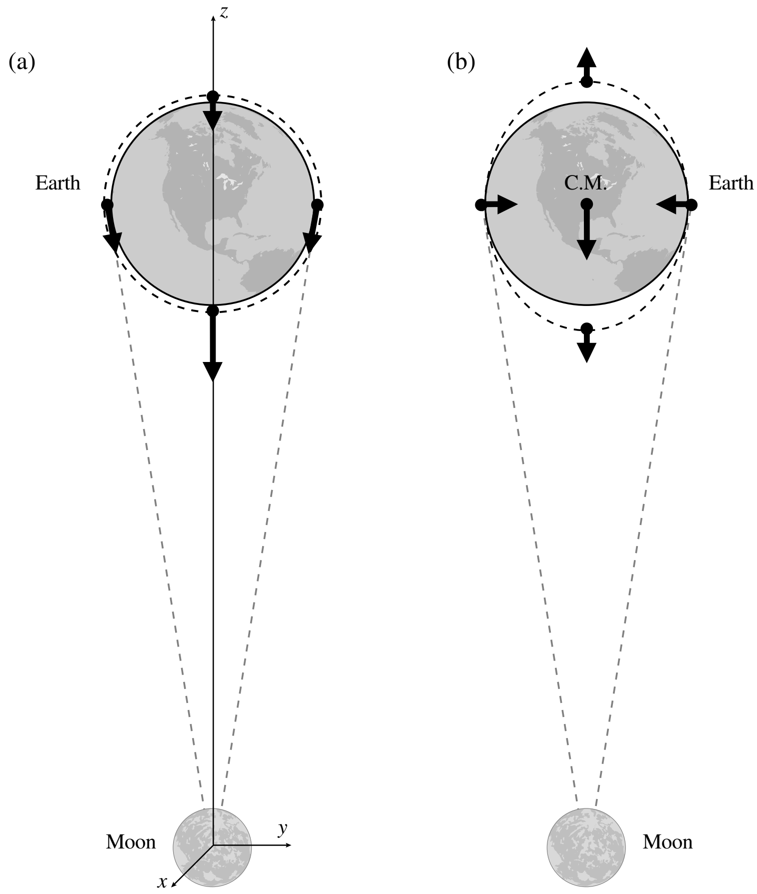

Fig. 5.5 Variations of the gravitational field as tidal forces. (a) Lunar gravitational forces on four representative points on earth. (b) After taking out the center-of-mass motion, the relative forces on earth are the tidal forces giving rise to longitudinal stretching and transverse compression.

Since tidal forces cannot be coordinate-transformed away, they should be regarded as the essence of gravitation. They are the variations of the gravitational field, hence the second derivatives of the gravitational potential. From the discussion in this chapter showing that relativistic gravitational potential being the metric, and that second derivative of the metric being the curvature, we see that Einstein gives gravity a direct geometric interpretation by identifying these tidal forces with the curvature of spacetime. A discussion of tidal forces in terms of the Newtonian deviation equation is given in Box 5.2.

Box 5.2 The equation of Newtonian deviation and its GR generalization

Here we provide a more quantitative description of the gravitational tidal force in the Newtonian framework, which will suggest an analogous GR approach to be followed in Chapter 12.

As the above discussion indicates, the tidal effect concerns the relative motion of particles in an nonuniform gravitational field. Let us consider two particles: one has the trajectory $\mathbf{x}(t)$ and another has $\mathbf{x}(t) + \mathbf{s}(t)$. Namely, these two particles measured at the same time have a separation distance of $\mathbf{s}(t)$. The respective equations of motion $(i = 1, 2, 3)$ obeyed by these two particles are:

$$\frac{d^2x^i}{dt^2} = -\frac{\partial\Phi(x)}{\partial x^i} \quad \text{and} \quad \frac{d^2x^i}{dt^2} + \frac{d^2s^i}{dt^2} = -\frac{\partial\Phi(x+s)}{\partial x^i}. \tag{5.30}$$

Consider the case where $s^i(t)$ is small and we can approximate the gravitational potential $\Phi(x+s)$ by a Taylor expansion

$$\Phi(x+s) = \Phi(x) + \frac{\partial\Phi}{\partial x^j}s^j + \cdots. \tag{5.31}$$

From the difference of the two equations in (5.30), we obtain the **Newtonian deviation equation** that describes the separation between two particle trajectories in a gravitational field

$$\frac{d^2s^i}{dt^2} = -\frac{\partial^2\Phi}{\partial x^i\partial x^j}s^j. \tag{5.32}$$

Thus the relative acceleration per unit separation is given by a tensor having the second derivatives of the gravitational potential (i.e. the tidal force components) as its elements.

We now apply (5.32) to the case of a spherical gravitational source, for example, the gravity due to the moon on earth, see Fig. 5.5,

$$\Phi(x) = -\frac{G_{\mathrm{N}}M}{r}, \tag{5.33}$$

where the radial distance is related to the rectangular coordinates by $r = (x^2+y^2+z^2)^{1/2}$. Since $\partial r/\partial x^i = x^i/r$ we have

$$\frac{\partial^2\Phi}{\partial x^i\partial x^j} = \frac{G_{\mathrm{N}}M}{r^3}\left(\delta_{ij} - \frac{3x^ix^j}{r^2}\right). \tag{5.34}$$

Consider the case of the "first particle" being located along the z-axis $x^i = (0, 0, r)$, the Newtonian deviation Eq. (5.32) for the displacement of the "second particle," with the second derivative tensor given by (5.34), now takes on the form of

$$\frac{d^2}{dt^2}\begin{pmatrix} s_x \\ s_y \\ s_z \end{pmatrix} = \frac{-G_{\mathrm{N}}M}{r^3}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix}\begin{pmatrix} s_x \\ s_y \\ s_z \end{pmatrix}. \tag{5.35}$$

We see that there is an attractive tidal force between the two particles in the transverse direction $\mathfrak{T}_{x,y} = -G_{\mathrm{N}}Mr^{-3}s_{x,y}$ that leads to compression; a tidal repulsion $\mathfrak{T}_z = +2G_{\mathrm{N}}Mr^{-3}s_z$, leading to stretching, in the longitudinal (i.e. radial) direction.

In GR we shall follow a similar approach (see Problems 12.4 and 12.5): the two equations of motion (5.30) will be replaced by the corresponding geodesic equations; their difference, after a Taylor expansion, leads to the **equation of geodesic deviation**, which is entirely similar[3] to (5.32). Since the metric function is the relativistic potential, the second derivative tensor turns into the curvature tensor of the spacetime (the Riemann curvature tensor). In this geometric language we see that the cause of the deviation from flat spacetime worldline is attributed to the curvature (cf. previous discussion in Section 5.1.2.)

[3] We are not quite ready to derive this GR equation as one still needs to learn (in Chapter 11) how to perform differentiations in a curved space.

5.3.2 The GR field equation described

We now show that the curvature, identified as the tidal forces, enters directly in the field equations of relativistic gravitational theory.

The field equation relates the source distribution to the resultant field. Namely, given the source distribution, we can use the field equation to find the field everywhere. For the Newtonian Eq. (5.5),

$$\nabla^2\Phi = 4\pi G_{\mathrm{N}}\rho, \tag{5.36}$$

we have the second derivative of the gravitational potential $\nabla^2\Phi$ being directly proportional to the mass density ρ. What is the relativistic generalization of this equation?

1. For the right-hand side (RHS) of the field equation, from the viewpoint of relativity, mass being just a form of energy (the rest energy) and, furthermore, energy and momentum being equivalent: they can be transformed into each other when viewed by different observers (cf. (2.62)), the mass density ρ of (5.5) is generalized in relativity to an object called the "**energy–momentum tensor**" $T_{\mu\nu}$. The 16 elements include 1 being the energy density $T_{00} = \rho c^2$, 3 elements T_{0i} the momentum densities, and the remaining 12 elements representing the fluxes associated with energy and momentum densities—they describe the flow of energy and momentum components. There are actually only 10 independent elements, because it is a symmetric tensor $T_{\mu\nu} = T_{\nu\mu}$. A more detailed discussion of the energy–momentum tensor will be presented in Section 10.4.
2. For the second derivative of the potential on the left-hand side (LHS) of the field equation, we have already seen that the relativistic gravitational potential is the metric $g_{\mu\nu}$ and the curvature in (4.35) is a second derivative of the metric. And, as we shall find in Chapter 11, for higher dimensional spaces, the Gaussian curvature K is generalized to the **Riemann curvature tensor**. A particular (contracted) version of the curvature tensor is the **Einstein tensor** $G_{\mu\nu}$, having mathematical properties that match those of the energy–momentum tensor $T_{\mu\nu}$.

This suggests the possible relativistic generalization of the gravitational field equation as having the basic structure of (5.36): the RHS being the energy–momentum tensor, and the LHS being the Einstein tensor involving the second derivative of the metric.

Newton	$\nabla^2\Phi \propto \rho$
GR	$G_{\mu\nu} \propto T_{\mu\nu}$

In this way, we obtain in Section 12.2 the **Einstein field equation** in the form of

$$G_{\mu\nu} = \kappa T_{\mu\nu}, \tag{5.37}$$

where κ is a proportionality constant. As we shall show in Section 12.2.2 the nonrelativistic limit of this equation is just the Newton's Eq. (5.36) when we make the identification of

$$\kappa = -\frac{8\pi G_{\mathrm{N}}}{c^4}. \tag{5.38}$$

Since the curvature has a different measurement unit from that for the energy–momentum density, the proportional constant κ, hence Newton's constant G_{N}, should be interpreted as a **conversion factor**. Just as the speed of light c is the conversion factor between space and time that is fundamental to the special relativistic symmetry of space and time (cf. Section 2.3), one way of viewing the significance of Newton's constant is that it is the conversion factor fundamental for a geometric description of gravity by GR, it connects spacetime curvature to the gravitational source of energy and momentum, as in Einstein equation:

$$\begin{pmatrix}\text{curvature}\\ \text{of spacetime}\end{pmatrix} = (\text{Newton's constant}) \times \begin{pmatrix}\text{energy–momentum}\\ \text{density}\end{pmatrix}.$$

When worked out in Chapter 12, we shall see that (5.37) represents 10 coupled partial differential equations. Their solution is the metric function $g_{\mu\nu}(x)$, fixing the geometry of spacetime. We emphasize once more that in GR, spacetime is no longer a passive background against which physical events take place. Rather, it is ever-changing as it responds to the ever-changing matter/energy distribution in the world.

For the rest of Parts I and II (Chapters 6–9), this Einstein field equation will not be discussed further. Rather, we shall concentrate on investigating its solutions, showing how a curved spacetime description of the gravitational field, that is, knowing the metric $g_{\mu\nu}(x)$, brings about many interesting physical consequences: from bending of light rays, black holes, to cosmology.

Box 5.3 Einstein's three motivations: an update

In Chapter 1 we discussed Einstein's motivations for creating GR. Now we can see how these issues are resolved in the curved spacetime formulation of a relativistic theory of gravitation.

1. **SR is not compatible with gravity.** In the GR formulation, we see that SR is valid only in the locally inertial frames in which gravity is transformed away.
2. **A deeper understanding of $m_{\mathrm{I}} = m_{\mathrm{G}}$.** The weak EP is generalized to strong EP. The various consequences of EP led Einstein to the idea of a curved spacetime as the relativistic gravitational field. At the fundamental level there is no difference between gravity and the "fictitious forces" associated with accelerated frames. Noninertial frames of reference in Newtonian physics are identified in Einstein's theory with the presence of gravity. The GR theory, symmetric with respect to general coordinate transformations (including accelerated coordinates), and relativistic field theory of gravitation must be one and the same. EP is built right into the curved spacetime description of gravitation because any curved space is locally flat.

3. **"Space is not a thing."** GR equations are covariant under the most general (position-dependent) coordinate transformations. (See further discussion as the "principle of general covariance" in Section 12.1.) Spacetime is the solution to the Einstein equation. It has no independent existence except expressing the relation among physical processes in the world.

Review questions

1. What does one mean by a "geometric theory of physics"? Use the distance measurements on the surface of earth to illustrate your answer.
2. How can the phenomenon of gravitational time dilation be phrased in geometric terms? Use this discussion to support the suggestion that the spacetime metric can be regarded as the relativistic gravitational potential.
3. Give a simple example how the EP physics implies a non-Euclidean geometric relation.
4. What significant conclusion did Einstein draw from the analogy between the fact that a curved space is locally flat and that gravity can be transformed away locally?
5. How does GR imply a concept of space and time as reflecting merely the relationship between physical events rather than a stage onto which physical events take place?
6. Give the heuristic argument for GR equation of motion to be the geodesic equation.
7. What is the Newtonian limit? In this limit, what relation can one infer between the Newtonian gravitational potential and a metric tensor component of the spacetime? Use this relation to derive the gravitational Doppler shift.
8. What are tidal forces? How are they related to the gravitational potential? Explain why the solar tidal forces are smaller than lunar tidal forces, even though the gravitational attraction of earth by the sun is stronger than that by the moon. Explain how in GR the tidal forces are identified with the curvature of spacetime.
9. Give a qualitative description of the GR field equation. Explain in what sense we can regard Newton's constant as a basic "conversion factor" in GR. Can you name two other conversion factors in physics that are basic respectively to SR and to quantum theory?
10. How are Einstein's three motivations for creating GR resolved in the final formulation of the geometric theory of gravity?

Problems

(5.1) **The metric element g_{00}** From the definitions of metric and propertime, derive the relation between proper time and coordinate time

$$d\tau = \sqrt{-g_{00}}dt.$$

(5.2) **Spatial distance and spacetime metric** Einstein suggested the following definition of spatial distance dl between two neighboring points (A, B) with a coordinate difference of dx^i (where $i = 1, 2, 3$): a light pulse is sent to B and reflected back to A. If the elapsed proper time (according to A) is $d\tau_A$, then $dl \equiv c d\tau_A/2$. The square of the spatial distance should also be quadratic in dx^i:

$$dl^2 = \gamma_{ij} dx^i dx^j.$$

How is this spatial metric related to the spacetime metric $g_{\mu\nu}$ as defined by $ds^2 = g_{\mu\nu}dx^\mu dx^\nu$ (where $\mu = 0, 1, 2, 3$)? Is it just $\gamma_{ij} = g_{ij}$? (cf. Landau and Lifshitz, 1975, §84).

(5.3) **Non-Euclidean geometry of a rotating cylinder** In Section 5.1.1 we used the example of a rotating cylinder to motivate the need of non-Euclidean geometry. Use the formalism derived in Problem 5.2 to work out the spatial distance, showing this violation of the Euclidean relation between radius and circumference.

(5.4) **Geodesic equation in a rotating coordinate** Knowing the metric for a rotating coordinate from Problem 5.3, work out the corresponding Christoffel symbol and geodesic equation. This can be taken as the relativistic version of the centrifugal force.

(5.5) **The geodesic equation and light deflection** Use the geodesic equation, rather than Huygens' principle, to derive the expression of gravitational angular deflection given by (3.44), if the only warped metric element is $g_{00} = -1 - 2\Phi(x)/c^2$. One approach is to note (see Fig. 3.6) that the infinitesimal angular deflection $d\delta$ of a photon with a momentum $\mathbf{p} = p\hat{\mathbf{x}}$ is related to momentum change by $d\delta = dp_y/p$. Momentum in turn is proportional to differentiation of displacement with respect to the proper time $p \propto dx/d\tau$ so we have the relation between the two 4-vectors: $p^\mu \propto dx^\mu/d\tau$ with $\mu = 0, 1, 2, 3$. For the Minkowski metric $\eta_{\mu\nu} = \text{diag}(-1, 1, 1, 1)$, a light-like ($\eta_{\mu\nu}dx^\mu dx^\nu = 0$) displacement along the x direction leads to $dx^\mu = (dx, dx, 0, 0)$, that is, $dx^0 = dx^1 = dx$ and the momentum 4-vector $p^\mu = (p, p, 0, 0)$, as appropriate for massless photon [cf. (2.62) and (2.64)]. The deflection can be calculated from the geodesic equation by its determination of $dx^\mu/d\tau$, hence p^μ.

(5.6) **Symmetry property of the Christoffel symbols** From the definition of (5.10), check explicitly that

$$\Gamma^\lambda_{\mu\nu} = \Gamma^\lambda_{\nu\mu}.$$

(5.7) **The matrix for tidal forces is traceless** One notes that the matrix in (5.35) is traceless (vanishing sum of the diagonal elements). Why should this be so?

(5.8) **G_N as a conversion factor** From Newton's theory we know that Newton's constant has the dimension of (energy) (length) (mass)$^{-2}$. With such a G_N in the proportional constant (5.38) of the Einstein Eq. (5.37), check that it yields the correct dimension for the curvature on the LHS of the Einstein equation. (NB the elements of the energy–momentum tensor have the dimension of energy density.)

Spacetime outside a spherical star

6

- Spacetime outside a spherical source has the *Schwarzschild geometry.*
- Gravitational lensing and the precession of Mercury's perihelion worked out as examples of geodesics in the Schwarzschild spacetime.
- *Black hole* is an object so compact that it is inside its *Schwarzschild surface*, which is an *event horizon*: an observer outside cannot receive any signal sent from inside.
- The physical reality of, and observational evidence for, black holes is briefly discussed.

In the previous chapter we presented some preliminaries for a geometric description of gravity. The gravitational field as curved spacetime can be expressed (once a coordinate system has been chosen) in terms of the metric function. In this chapter, we discuss the solution $g_{\mu\nu}(x)$ to the Einstein field equation for the region outside a spherically symmetric source (e.g. the sun), called the (exterior) **Schwarzschild geometry**. Its nonrelativistic analog is the gravitational potential

$$\Phi(r) = -\frac{G_{\mathrm{N}}M}{r}, \tag{6.1}$$

which is the solution to the Newtonian field equation $\nabla^2\Phi\ (x) = 4\pi G_{\mathrm{N}}\rho(x)$ with a spherically symmetric mass density $\rho(x)$ and a total mass M inside a sphere with radius less than the radial distance r.

The general relativity (GR) field equation with a spherical source will be solved in Chapter 12. Given the source mass distribution, we can find the spacetime $g_{\mu\nu}(x)$ outside a spherical star. In this chapter, we shall only quote the solution, called the Schwarzschild metric, and concentrate on the description of a test particle's motion in this geometry. In this connection we study several interesting applications: gravitational lensing, precession of planet Mercury's perihelion, and finally, (nonrotating) black holes.

6.1 Description of Schwarzschild spacetime

We shall first show that, in a spherical coordinate system (t, r, θ, ϕ), the metric tensor for a spherically symmetric spacetime has only two unknown elements. These scalar metric functions $g_{00}(t, r)$ and $g_{rr}(t, r)$ can be obtained by solving the GR field equation. This explicit solution will not be carried out until Section 12.3. Here we just present the result:

$$g_{00}(t, r) = -\frac{1}{g_{rr}(t, r)} = -1 + \frac{2G_{\mathrm{N}}M}{rc^2}. \tag{6.2}$$

The implication of such a geometry for various physical situations will then be discussed in subsequent sections.

6.1.1 Spherically symmetric metric tensor

Because the source is spherically symmetric, the spacetime it generates (as the solution to the Einstein equation for such a source) must have this symmetry. The corresponding metric function $g_{\mu\nu}(x)$ must be isotropic in the spatial coordinates. It is natural to pick a spherical coordinate system (t, r, θ, ϕ) having the center coincident with that of the spherical source. As shown in Box 6.1, the form of such an isotropic metric has only two unknown scalar functions g_{00} and g_{rr}:

$$g_{\mu\nu} = \mathrm{diag}(g_{00}, g_{rr}, r^2, r^2 \sin^2\theta) \tag{6.3}$$

which, when far away from the gravitational source, approaches the flat spacetime limit:

$$\lim_{r\to\infty} g_{00}(t, r) \to -1 \quad \text{and} \quad \lim_{r\to\infty} g_{rr}(t, r) \to 1. \tag{6.4}$$

Box 6.1 The standard form of an isotropic metric

Here we shall explicitly show that a spherically symmetric metric tensor has only two unknown scalar functions.

1. **General considerations of isotropy.** The infinitesimal invariant interval $ds^2 = g_{\mu\nu}dx^\mu dx^\nu$ must be quadratic in $d\mathbf{x}$ and dt without singling out any particular spatial direction. Namely, ds^2 must be composed of terms having two powers of $d\mathbf{x}$ and/or dt; the vectors $\mathbf{x}$ and $d\mathbf{x}$ must appear in the form of dot products so as not to spoil the spherical symmetry. The vector $\mathbf{x}$ can appear because the metric is a function of position and time $g_{\mu\nu} = g_{\mu\nu}(\mathbf{x}, t)$.

$$ds^2 = Ad\mathbf{x}\cdot d\mathbf{x} + B(\mathbf{x}\cdot d\mathbf{x})^2 + Cdt(\mathbf{x}\cdot d\mathbf{x}) + Ddt^2, \tag{6.5}$$

where $A, B, C,$ and D are scalar functions of t and $\mathbf{x}\cdot\mathbf{x}$. In a spherical coordinate system (r, θ, ϕ):

$$\mathbf{x} = r\hat{\mathbf{r}}, \quad \text{and} \quad d\mathbf{x} = dr\hat{\mathbf{r}} + rd\theta\hat{\boldsymbol{\theta}} + r\sin\theta d\phi\hat{\boldsymbol{\phi}}. \tag{6.6}$$

Thus

$$\mathbf{x}\cdot\mathbf{x} = r^2 \quad \text{and} \quad \mathbf{x}\cdot d\mathbf{x} = rdr,$$
$$d\mathbf{x}\cdot d\mathbf{x} = dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2).$$

The invariant separation written in terms of spherical coordinates is now

$$ds^2 = A[dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2)] + Br^2dr^2 + Crdrdt + Ddt^2, \tag{6.7}$$

or, equivalently, with some relabeling of scalar functions,

$$ds^2 = A[r^2(d\theta^2 + \sin^2\theta d\phi^2)] + Bdr^2 + Cdrdt + Ddt^2. \tag{6.8}$$

2. **Simplification by coordinate choices.** From our discussion in Chapter 4, we learnt that the Gaussian coordinates, as labels of points in

space, can be freely chosen. In the same way, the name given to coordinates in Riemannian geometry has no intrinsic significance until their connection to physical length ds^2 is specified by the metric function. Thus, we are free to make new choices of coordinates (with corresponding modification of the metric) until the metric takes on the simplest form. Of course, in our particular case, the process of changing to new coordinates must not violate spherical symmetry.

(a) **New time coordinate so that there will be no cross $dtdr$ term.** Introducing a new coordinate t'

$$t \rightarrow t' = t + f(r). \tag{6.9}$$

We have $dt' = dt + (df/dr)dr$ and

$$dt^2 = dt'^2 - \left(\frac{df}{dr}\right)^2 dr^2 - 2\frac{df}{dr}drdt. \tag{6.10}$$

Now the cross $dtdr$ term has a coefficient $C - 2D(df/dr)$, which can be eliminated by choosing an $f(r)$ that satisfies the differential equation

$$\frac{df}{dr} = +\frac{C}{2D}. \tag{6.11}$$

Incidentally, the absence of any linear dt factor means that the metric is also time-reversal invariant.

(b) **New radial coordinate so that the angular coefficient is trivial.** We can set the function A in (6.8) to unity by choosing a new radial coordinate

$$r^2 \rightarrow r'^2 = A(r,t)r^2$$

so that the first term on the right-hand side (RHS) of (6.8) is just $r'^2(d\theta^2 + \sin^2\theta d\phi^2)$.

With these new coordinates we are left with only two unknown scalar functions in the metric. In this way, the interval takes on the form of

$$ds^2 = g_{00}(r,t)c^2dt^2 + g_{rr}(r,t)dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2), \tag{6.12}$$

which is shown in (6.3).

Interpreting the coordinates

Our spherical symmetric metric (6.3) is diagonal. In particular, $g_{i0} = g_{0i} = 0$ (with i being a spatial coordinate index) means that for a given t, we can discuss the spatial subspace separately. For a fixed t, we can visualize this spherically symmetric coordinate system as a series of 2-spheres having different radial coordinate values of r, with their center at the origin of the spherically symmetric source. Each 2-sphere, having surface area of $4\pi r^2$ and volume $4\pi r^3/3$, can be thought of as made up of rigid rods arranged in a grid corresponding to various (θ, ϕ) values with synchronized clocks attached at each grid point (i.e. each point in this subspace has the same coordinate time).

- Before the gravitational source is "turned on," the coordinate r is the proper radial distance ρ defined as

$$dr^2 = ds_r^2 \equiv d\rho^2, \tag{6.13}$$

where ds_r^2 is the invariant interval ds^2 with $dt = d\theta = d\phi = 0$, and the coordinate t is the proper time τ for an observer at a fixed location

$$-c^2dt^2 = ds_t^2 \equiv -c^2d\tau^2, \tag{6.14}$$

where ds_t^2 is the invariant interval ds^2 with $dr = d\theta = d\phi = 0$. Thus the coordinates $(r,\ t)$ have the physical interpretation as the radial distance and time measured by an observer far away from the (spherical) gravitational source.

- After "turning on" the gravitational source, we have a warped spacetime. In particular $g_{rr} \neq 1$ there is curving in the spatial radial direction so that the proper radial distance $\rho \neq r$ as

$$d\rho = \sqrt{g_{rr}}dr. \tag{6.15}$$

Consequently, the spherical surface area $4\pi r^2$ and volume $4\pi r^3/3$ no longer bear the Euclidean relation with their radius ρ. Similarly, the proper time

$$d\tau = \sqrt{-g_{00}}dt, \tag{6.16}$$

differs from the coordinate time because $g_{00} \neq -1$. It signifies the warping of the spacetime in the time direction.

6.1.2 Schwarzschild geometry

In GR, spacetime is not a passive background against which physical processes take place. Rather, the geometry is determined by the distribution of mass and energy. Given the source distribution, we can solve the GR field equation to find the metric $g_{\mu\nu}(x)$. The solution of Einstein's equation for the spacetime **exterior** to a spherical source will be carried out in Chapter 12. Here we quote this result

$$g_{00}(t,r) = -\frac{1}{g_{rr}(t,r)} = -1 + \frac{r^*}{r}, \tag{6.17}$$

where r^* is some constant length scale. We see that the deviation from flat spacetime of (6.4) is determined by the size of the ratio r^*/r. The resultant metric is called the **Schwarzschild metric**,

$$g_{\mu\nu} = \text{diag}\left[\left(-1 + \frac{r^*}{r}\right), \left(1 - \frac{r^*}{r}\right)^{-1}, r^2, r^2\sin^2\theta\right]. \tag{6.18}$$

The Schwarzschild radius

We can relate this constant distance r^* to familiar quantities by considering the Newtonian limit (5.20) and (6.1) that lead to

$$g_{00} = -\left(1 + \frac{2\Phi}{c^2}\right) = -1 + \frac{2G_\text{N}M}{c^2r}. \tag{6.19}$$

Comparing this to (6.17), we obtain the **Schwarzschild radius**

$$r^* = \frac{2G_\text{N}M}{c^2}. \tag{6.20}$$

It is generally a very small distance: for example, the solar and terrestrial Schwarzschild radii are respectively:

$$r^*_\odot \simeq 3\,\mathrm{km} \quad \text{and} \quad r^*_\oplus \simeq 9\,\mathrm{mm}. \tag{6.21}$$

Hence, in general, the ratio r^*/r, which signifies the modification of the flat Minkowski metric, is a very small quantity. For the exterior solutions to be applicable, the smallest value that r can take is the radius R of the spherical source: the above r^* values translate into

$$\frac{r^*_\odot}{R_\odot} = O\,(10^{-6}) \quad \text{and} \quad \frac{r^*_\oplus}{R_\oplus} = O\,(10^{-10}). \tag{6.22}$$

The metric (6.18) is singular at the Schwarzschild radius. This singular feature was not extensively studied at the beginning because many early relativists thought $r = r^*$ was not physically realizable. Because r^* is so small and for the exterior solution $r > r_{\text{source}}$ to be applicable, the massive source must have such an extraordinary density that $r_{\text{source}} < r^*$. Only gradually was such a possibility taken seriously. This situation of the black holes will be discussed in Section 6.4. Our discussion here suggests that a black hole is expected to have extremely high mass density. In fact a stellar-mass black hole $M = O\,(M_\odot)$ indeed has a high density $O\,(10^{19}\,\mathrm{kg/m^3})$, comparable to nuclear density. But we must keep in mind that black hole density[1] $(r^*)^{-3}M$ actually has an inverse dependence on it mass $\propto M^{-2}$ because the Schwarzschild radius r^* increases with mass. Thus for supermassive galactic back hole $M = O\,(10^9 M_\odot)$ the density is less than water!

Embedding diagram

A helpful way to visualize the warped space is to use an **embedding diagram**. Since it is difficult to work with the full 3D curved space, we shall concentrate on the 2D subspace, corresponding to a fixed polar angle $\theta = \pi/2$ (and at some given instance of time). Namely, we will focus on the 2D space slicing across the middle of the source. In the absence of gravity this is just a flat plane as depicted in Fig. 6.1(a). In the presence of gravity, this is a curved 2D space. We would like to have a way to visualize the warped nature of this 2D subspace (outside the source). A helpful way to do this is to imagine that this curved 2D surface is embedded in a fictitious 3D Euclidean space, Fig. 6.1(b).

A particle moving on this $\theta = \pi/2$ plane will naturally trace out a bent trajectory as it follows the geodesic of this warped surface in the embedding space. In our illustration we have used the Schwarzschild solution for a compact source

$$\Delta\rho = \Delta r \left(1 - \frac{r^*}{r}\right)^{-1/2}. \tag{6.23}$$

In this embedding diagram, the distance from the center on the curved surface is ρ while that in the horizontal plane is the coordinate radial distance r. Thus a small change in r corresponds to a large change in ρ when r approaches r^*.

[1] Based on the fact that no measurements can be made inside a black hole, we chose to define the density of a black hole as the ratio of its mass to the spherical volume with radius r^*.

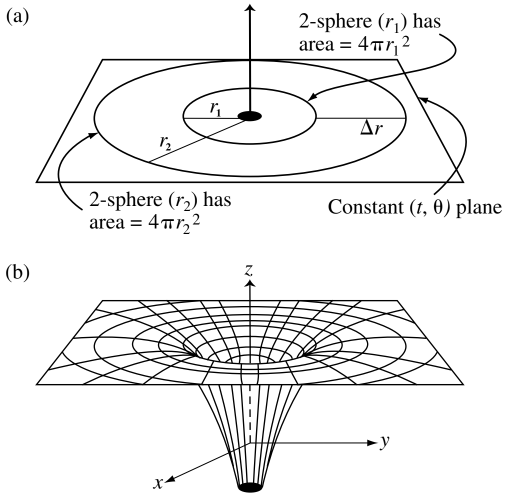
(a)
2-sphere (r_1) has area = $4\pi r_1{}^2$
r_1
r_2
Δr
2-sphere (r_2) has area = $4\pi r_2{}^2$
Constant (t, θ) plane
(b)
z
y
x

Fig. 6.1 (a) The $\theta = \pi/2$ plane (r, ϕ) cutting across the spherical source. (b) In a fictitious 3D embedding space, the physical 2D subspace of (a) is shown as a curved surface. In this example, we have used the Schwarzschild solution of $g_{rr} = (1 - r^*/r)^{-1}$. The singular nature of the space at $r \simeq r^*$ is reflected by the steep slope of the curved surface near the Schwarzschild circle.

Isotropic metric is time-independent

We note from (6.17) that the scalar metric functions are time-independent even though we have not assumed a constant source. This turns out to be a general result (**Birkhoff theorem**): whenever the source is isotropic the resultant spacetime must necessarily be time-independent. The theorem will be proven in Box 12.3. In the meantime it is worthwhile to point out that the same result holds for the Newtonian theory as well. Recall that the gravitational field outside a spherical source is identical to the gravitational field due to a point source having all the spherical mass at the center of the spherical source. This proof depends only on the symmetry property of the problem, and is not affected by any possible time dependences. Thus, regardless of whether the spherical mass is pulsating or exploding, etc., the resultant field is the same, as long as the spherical symmetry is maintained. The analogous situation in electromagnetism is the statement that there is only dipole, quadrupole, ..., but no monopole, radiation.

6.2 Gravitational lensing

From the consideration of the equivalence principle (EP), Einstein already deduced (see Section 3.3.2) that there will be a bending of the star light grazing past the sun. This effect is closely related to the idea of gravitational time dilation, expressed as a deviation from the flat space metric, see (5.3),

$$g_{00} = -\left(1 + \frac{2\Phi}{c^2}\right) = -1 + \frac{2G_{\mathrm{N}}M}{c^2 r} = -1 + \frac{r^*}{r}, \qquad (6.24)$$

which we see is part of the exact Schwarzschild solution (6.18). In the full GR theory, the warping of the spacetime takes place not only in the time direction $g_{00} \neq -1$, but in the radial spatial direction as well, $g_{rr} \neq 1$. Here we calculate the effect of this extra warping on the bending of the light-ray, finding a doubling of the deflection angle. The bending of the light ray by a massive object can be linked to that by a lens. In Section 6.2.2 we shall present the lens equation, and discuss gravitational lensing as an important tool for modern astronomy.

6.2.1 Light ray deflection revisited

Let us consider the lightlike worldline in a fixed direction $d\theta = d\phi = 0$,

$$ds^2 = g_{00}c^2dt^2 + g_{rr}dr^2 = 0. \tag{6.25}$$

To an observer far from the source, using the coordinate time and radial distance (t, r), the effective light speed according to (6.25) is

$$c(r) \equiv \frac{dr}{dt} = c\sqrt{-\frac{g_{00}(r)}{g_{rr}(r)}}. \tag{6.26}$$

A slightly different way of arriving at the same result is by noting that the speed of light is absolute in terms of physical quantities $c = d\rho/d\tau$ which is just (6.26), after using (6.15) and (6.16) which relate the proper distance and time (ρ, τ) to the coordinate distance and time (r, t).

In the previous EP discussion, we had effectively set $g_{rr} = 1$. Now the Schwarzschild solution (6.17) informs us that $g_{00} = -g_{rr}^{-1}$. The influence of $g_{00} \neq -1$ and $g_{rr} \neq 1$ in (6.26) are of the same size and in the same direction. Thus the deviation of the vacuum index of refraction $n(r)$ from unity is **twice** as large as that when only the EP effect was taken into account as in (3.40):

$$n(r) = \frac{c}{c(r)} = \sqrt{-\frac{g_{rr}(r)}{g_{00}(r)}} = \frac{1}{-g_{00}(r)} = 1 - 2\frac{\Phi(r)}{c^2}. \tag{6.27}$$

Namely, the retardation of a light signal is twice as large as that given in (3.39)

$$c(r) = \left(1 + 2\frac{\Phi(r)}{c^2}\right)c. \tag{6.28}$$

According to Eqs (3.39)–(3.45) the deflection angle δ, being directly proportional to this deviation, is twice as large as that given by (3.47) (see also Problem 5.2):

$$\delta_{\mathrm{GR}} = 2\delta_{\mathrm{EP}} = \frac{4G_{\mathrm{N}}M_{\odot}}{c^2 r_{\mathrm{min}}}. \tag{6.29}$$

We should apply $r_{\mathrm{min}} = R_{\odot}$ for the case of the light ray grazing past the sun.

This predicted deflection of 1.74 arc seconds (about 1/4000 of the angular width of the sun as seen from earth) is not easy to detect. One needed a solar eclipse against the background of several bright stars (so that some could be used as reference points). The angular position of a star with light grazing past the (eclipsed) sun would appear to have moved to a different position when compared to the location in the absence of the sun (cf. Fig. 3.6). On May 29, 1919 there was such an eclipse. Two British expeditions were mounted: one to Sobral in northern Brazil, and another to the island of Principe, off the coast of West Africa. The report by Dyson, Eddington, and Davidson that Einstein's prediction was successful in these tests created a worldwide sensation, partly for scientific reasons, and partly because the world was amazed that so soon after the First World War the British should finance and conduct an expedition to test a theory proposed by a German citizen.

6.2.2 The lens equation

Gravitational deflection of a light ray discussed above, Fig. 6.2(a), has some similarity to the bending of light by a glass lens, Fig. 6.2(b). The difference is

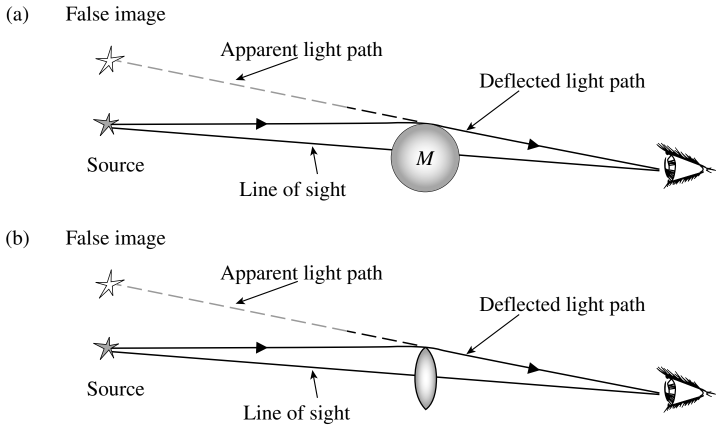

Fig. 6.2 Bending of light ray as a lensing effect. (a) Light from a distant star (source) is deflected by a lensing mass M lying close to the line of sight from the observer to the source. As a result, the source star appears to be located at a different direction. (b) Bending of light in (a) is analogous to that by a glass lens.

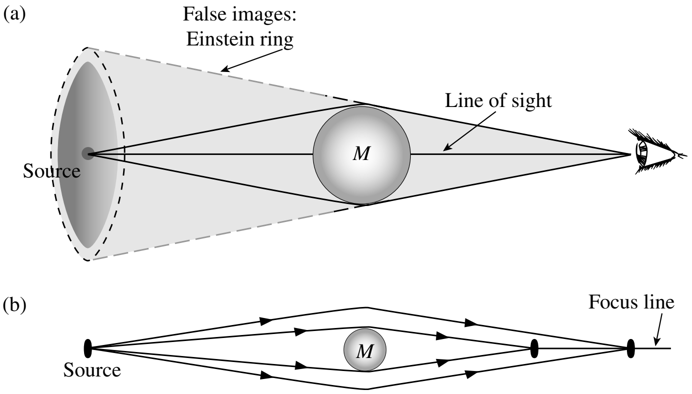

Fig. 6.3 Gravitational lensing of distant stars. (a) When the source and lensing mass are sufficiently far, double images can result. If the line of sight passes directly through the center of symmetrical lensing mass distribution, the false image appears as a ring, the Einstein ring. (b) Gravitational lens "focuses" images on a line.

that while a convex lens can focus images at a point, gravitational lenses "focus" on a line, Fig. 6.3(b). This is so because the gravitational deflection angle is a decreasing function of the impact parameter $b = r_{\min}$, as seen in (6.29), for a light ray passing through the lens the deflection angle is an increasing function. When the source and observers are sufficiently far from the lensing mass, bending from both sides of the lensing mass can produce double images, Fig. 6.3(a), and even a ring image. Solar deflection has been shown in our previous discussion to yield a bending angle of $\delta = 1.74''$. Such a small deflection means that the earth is too close to the sun[2] to see such lensing features as multiple images of background stars by the sun. The minimum distance needed is $R_\odot/(1.74''$ in radian$) \simeq 0.8 \times 10^{14}$ m $\gtrsim 500$ AU.

To the extent that we can approximate the light trajectory by its asymptotes,[3] we can derive the lens equation by simple geometrical consideration of Fig. 6.4(b). The distance SS$'$ can be obtained in two ways[4]: (i) It subtends the angle δ to yield SS$' = D_{\mathrm{ls}}\delta$, where D_{ls} is the distance between the lensing mass and the source light, and (ii) it is the length difference SS$'$ = S$'$O$'-$SO$'$ = $D_{\mathrm{s}}(\theta-\beta)$, where D_{s} is distance from observer to the source light, while $(\theta-\beta)$ is the angular separations between the image and source points. Equating these two expressions and plugging in the previous result of (6.29) with $b = r_{\min}$:

$$\delta = \frac{D_{\mathrm{s}}}{D_{\mathrm{ls}}}(\theta-\beta) = \frac{4G_{\mathrm{N}}M}{bc^2}, \tag{6.30}$$

[2] One astronomical unit (AU) $= 1.5\times10^{13}$ cm is the mean distance from earth to the sun.

[3] This is the **thin lens approximation**: all the action of the deflection is assumed to take place at one position. It is valid only if the relative velocities of the lens, source, and observer are small compared to the velocity of light, $v \ll c$.

[4] Here we concentrate on S$' \equiv$ S$'_+$. The calculation for S$'_-$ is the same.

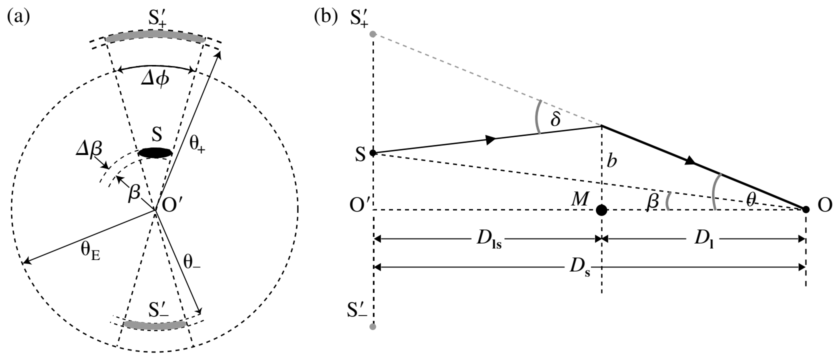

Fig. 6.4 Geometry of gravitational lensing. (a) Azimuthal and polar angle labels. (b) The observer and source are at O and S, respectively. The light ray is deflected by an angle of δ by the lensing mass M. The true and apparent positions S and S′ of the distant star are located by angles β and θ, respectively. b is the impact parameter $b = r_{\text{min}}$.

and approximating the impact parameter $b \approx D_l\theta$, we then have the **lens equation**

$$\beta = \theta - \frac{D_{ls}}{D_s D_l}\frac{4G_N M}{\theta c^2}. \tag{6.31}$$

This equation for a point lensing mass, being quadratic in its variable θ, usually has two solutions corresponding to the two images $S'_\pm$, resulting from bending on two sides of the lensing mass, Fig. 6.3.

In the special case of $\beta = 0$, that is, perfect alignment of the source, lens, and observer, the azimuthal axial symmetry of the problem yields a ring image, the **Einstein ring**, with angular radius

$$\theta_E = \sqrt{\frac{D_{ls}}{D_s D_l}\frac{4G_N M}{c^2}}. \tag{6.32}$$

For the general case when the source is not exactly behind the lens, the lens equation (6.31) for a single point lens

$$\beta = \theta - \theta_E^2/\theta \tag{6.33}$$

has two solutions:

$$\theta_\pm = \frac{1}{2}\left(\beta \pm \sqrt{\beta^2 + 4\theta_E^2}\right), \tag{6.34}$$

corresponding to presence of double images. One is inside the would-be Einstein ring $\theta_- < \theta_E$, the other outside. With the azimuthal angular width $\Delta\phi$ of the source unchanged, these two images are distorted into arcs, Fig. 6.4(a). Furthermore, if the distance to the source D_s and to the lens D_l can be estimated, the mass of the lens can be deduced by measurements of $\theta_\pm$ (hence θ_E) via (6.32).

Since 1979, astronomers have discovered several dozens of double quasar pairs: two quasars[5] having the same properties but separated by the few arcseconds. To produce such a sizable separation, the lensing mass is expected to be a galaxy (having billions of the solar mass). Even more dramatically, if the lens is not a single galaxy but an entire cluster of galaxies, the images can be clusters of distorted arcs. In Fig. 6.5 we display the distorted images of distant galaxies as lensed by the cluster Abell 2218.

[5] Quasars (quasi-stellar objects) are very luminous sources of small angular size at great cosmic distances. See further discussion in p110, Section 6.4.5.

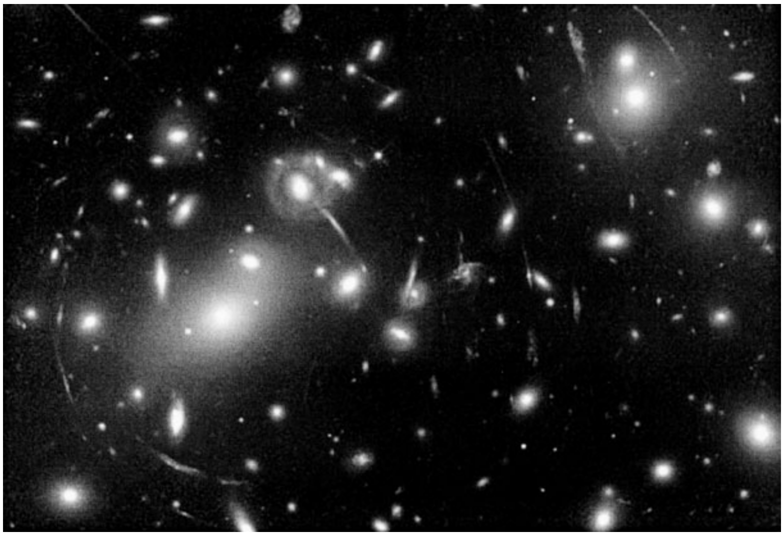

Fig. 6.5 Gravitational lensing effects due to the galaxy cluster Abell 2218. Just about all the bright objects in this picture taken by Hubble Space Telescope are galaxies in this cluster, which is so massive and so compact that it lenses the light from galaxies that lie behind it into multiple images of long faint arcs.

Box 6.2 Microlensing and the search for MACHOs

Gravitational lensing by stellar objects is typically too small to produce multiple images (i.e. the separate images cannot be resolved). Such **microlensing** events show up, because of the overlap of images, as an increase of luminosity flux of lensed sources. For the point lens discussed here we can calculate the magnification factor by noting that the light intensity (flux per unit solid angle of the source/image) is the same for each of the images $I = I_+ = I_-$ because they have the same source energy. Thus their flux is proportional to their respective subtended solid angles Ω, and the magnification is the ratio of combined image to that of the original flux in the absence of lensing mass:

$$\mu = \frac{f_+ + f_-}{f} = \frac{I_+ d\Omega_+ + I_- d\Omega_-}{I d\Omega} = \frac{\theta_+ d\theta_+ + \theta_- d\theta_-}{\beta d\beta}, \tag{6.35}$$

where we have used $d\Omega = \sin\theta d\theta d\phi \simeq \theta d\theta d\phi$ and the fact that azimuthal angular width is unchanged, $d\phi = d\phi_+ = d\phi_-$. We can calculate the individual magnification (either + or − image) by a simple differentiation of (6.33):

$$\frac{\theta d\theta}{\beta d\beta} = \left[1 - \left(\frac{\theta_{\mathrm{E}}}{\theta}\right)^4\right]^{-1}. \tag{6.36}$$

Plugging in the $\theta = \theta_\pm$ solutions of (6.34), we then obtain (Problem 6.3), for $\hat{\beta} = \beta/\theta_{\mathrm{E}}$, the magnification:

$$\mu = \frac{\hat{\beta}^2 + 2}{\hat{\beta}\sqrt{\hat{\beta}^2 + 4}} > 1. \tag{6.37}$$

Especially when $\hat{\beta} \to 0$, the magnification $\mu \propto 1/\hat{\beta}$ due to the whole Einstein ring-image can be quite significant. As an example of the gravitational lensing being a powerful tool of modern astronomy, evidence for the existence of "massive compact halo objects," or MACHOs has been obtained this way. As we shall discuss in the next chapter, there are compelling reasons to think that there are dark matter present in the outer reaches (the halo) of our galaxy. Some of this might be in the form of dead stars, black holes, or whatever massive objects that do not shine—such (baryonic) dark matter are collectively called the MACHOs (cf. Section 7.1.4). If a MACHO drifts in front of a background star, it will act as a lensing mass, thus enhancing the brightness of that star temporarily. Several astronomical teams undertook this search by simultaneously monitoring millions of stars in the Large Magellanic Cloud (a small satellite galaxy of the Milky Way). In 1997, the discovery of several such events (lasting a few weeks to several months) was announced (Alcock *et al.*, 1997).

6.3 Precession of Mercury's perihelion

In this section, we shall discuss the motion of a test mass in the Schwarzschild spacetime. In particular, we shall calculate the deviation from its Newtonian elliptical orbit.

Celestial mechanics based on Newtonian theory of gravitation has been remarkably successful. However, it had been realized around 1850 that there was a discrepancy between the theory and the observed **precession of the perihelion of the planet Mercury**. The pure $1/r^2$ force law of Newton predicts a closed elliptical orbit for a planet, that is, orbit with an axis **fixed** in space. However, the perturbations due to the presence of other planets and astronomical objects lead to a trajectory that is no longer closed. Since the perturbation is small, such a deviation from the closed orbit can be described as an ellipse with a **precessing** axis, Fig. 6.6. For the case of the Mercury, this planetary perturbation can account for most of the observed perihelion advances—5600″ (=1.556°) per century. However, there was still the discrepancy of 43 arc-second/century left unaccounted for. Following a similar situation involving Uranus that eventually led to the prediction and discovery of the outer planet Neptune in 1846, a new planet, named Vulcan, was predicted to lie inside the Mercury orbit. But it was never found. This is the perihelion precession problem that Einstein solved by applying his new theory of gravitation. GR implies a small correction to the $1/r^2$ force law, which just accounts for the missing 43″ advance of Mercury's orbit.

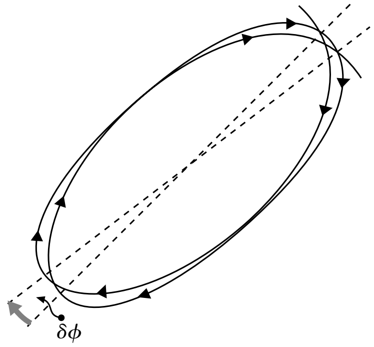

Fig. 6.6 A perturbed $1/r^2$ attraction leads to an open elliptical orbit which may be described as an elliptical orbit with a precessing axis. For planetary motion, this is usually stated as the precession of the minimal-distance point from the sun, the perihelion.

The GR problem we need to solve is as follows: given the gravitational field (the Schwarzschild spacetime due to the sun), we are to find the motion of the test particle (the Mercury planet). The geodesic equation that we need to solve is the Euler–Lagrange equation

$$\frac{\partial L}{\partial x^\mu} = \frac{d}{d\tau}\frac{\partial L}{\partial \dot{x}^\mu} \tag{6.38}$$

with the Lagrangian

$$L = \left(\frac{ds}{d\tau}\right)^2 = g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu. \tag{6.39}$$

$g_{\mu\nu}$ is the Schwarzschild metric, and, as is appropriate for a massive test particle, we have picked its proper time τ as the curve parameter and used the notation $\dot{x}^\mu = dx^\mu/d\tau$.

Just as in Newtonian mechanics, the trajectory (because of angular momentum conservation in a central force problem) will always remain in the plane spanned by the particle initial velocity and the vector **r** connecting the force center to the test particle. By setting $\theta = \pi/2$, the Lagrangian takes on the simple form of

$$L = -\left(1 - \frac{r^*}{r}\right) c^2 \dot{t}^2 + \left(1 - \frac{r^*}{r}\right)^{-1} \dot{r}^2 + r^2 \dot{\phi}^2 = -c^2. \tag{6.40}$$

The last equality follows from $L = (ds/d\tau)^2 = -c^2$ because $ds^2 = -c^2 d\tau^2$. NB only for a massive particle do we have $L = -c^2$; it is $L = 0$ for massless particles (cf. discussion of 4-velocity in Section 10.2).

The $\dot{\phi}$ and $\dot{t}$ terms are two constants of motion related to orbital angular momentum l and total (Newtonian) energy $\mathcal{K}$. Following the steps worked out in Box 6.3, (6.40) can be written as

$$\frac{1}{2} m \dot{r}^2 + \left(1 - \frac{r^*}{r}\right) \frac{l^2}{2mr^2} - \frac{G_N m M}{r} = \mathcal{K}. \tag{6.41}$$

Except for the $(1 - (r^*/r))$ factor, it is just the energy balance equation for the nonrelativistic central force problem. The extra factor of

$$-\frac{r^*}{r} \frac{l^2}{2mr^2} = -\frac{G_N M l^2}{mc^2 r^3} \tag{6.42}$$

may be regarded, for the problem of Mercury's orbit, as a small correction to the Newtonian potential energy $-G_N m M r^{-1}$ due to a r^{-4} type of force.

Box 6.3 Two constants of motion and the effective potential

Recall that if the Lagrangian $L = L(q, \dot{q})$ does not depend explicitly on one of the generalized coordinates q (so that $\partial L/\partial q = 0$), the Euler–Lagrangian equation implies the conservation law:

$$\frac{d}{d\tau}\left(\frac{\partial L}{\partial \dot{q}}\right) = 0. \tag{6.43}$$

In our case L does not explicitly depend on ϕ and t. The two corresponding constants of motion are essentially the orbital angular momentum (l) and the energy ($\mathcal{K}$).

$$\left(\frac{\partial L}{\partial \dot{\phi}}\right) = 2r^2 \dot{\phi} \equiv \lambda \tag{6.44}$$

and

$$\left(\frac{\partial L}{\partial \dot{t}}\right) = -2\left(1 - \frac{r^*}{r}\right) c^2 \dot{t} \equiv -2c^2 \eta. \tag{6.45}$$

After multiplying by $\frac{1}{2} m(1 - (r^*/r))$, and plugging in the constants λ and η, we can express the $L = -c^2$ Eq. (6.40) as

$$-\frac{mc^2 \eta^2}{2} + \frac{1}{2} m \dot{r}^2 + \left(1 - \frac{r^*}{r}\right) \frac{m\lambda^2}{8r^2} = -\frac{1}{2} mc^2 \left(1 - \frac{2G_N M}{c^2 r}\right) \tag{6.46}$$

or

$$\frac{1}{2} m \dot{r}^2 + \left(1 - \frac{r^*}{r}\right) \frac{m\lambda^2}{8r^2} - \frac{G_N m M}{r} = \frac{mc^2 \eta^2}{2} - \frac{1}{2} mc^2. \tag{6.47}$$

Renaming the constants according to

$$\frac{\lambda^2}{4} \equiv \frac{l^2}{m^2} \tag{6.48}$$

and

$$\frac{\eta^2 - 1}{2} \equiv \frac{\mathcal{K}}{mc^2} \tag{6.49}$$

this equation takes on the form of (6.41):

$$\frac{1}{2}m\dot{r}^2 + \left(1 - \frac{r^*}{r}\right)\frac{l^2}{2mr^2} - \frac{G_\mathrm{N}mM}{r} = \mathcal{K}. \tag{6.50}$$

This suggests that $\mathcal{K}$ has the interpretation of total (Newtonian) energy $\mathcal{K} = E - mc^2$, since the above equation is the energy balance equation:

$$\mathcal{K} = \frac{1}{2}m\dot{r}^2 + m\Phi_\mathrm{eff}, \tag{6.51}$$

with the effective potential being

$$\Phi_\mathrm{eff} = -\frac{G_\mathrm{N}M}{r} + \frac{l^2}{2m^2r^2} - \frac{r^*l^2}{2m^2r^3}. \tag{6.52}$$

This relativistic energy Eq. (6.41) can be cast in the form of an orbit equation. We can solve for $r(\phi)$ by the standard perturbation theory (see Box 6.4). With e being the eccentricity of the orbit, $\alpha = l^2/G_\mathrm{N}Mm^2 = (1+e)r_\mathrm{min}$ and $\epsilon = 3r^*/2\alpha$, the solution is

$$r = \frac{\alpha}{1 + e\cos[(1-\epsilon)\phi]}. \tag{6.53}$$

Thus the planet returns to its perihelion r_min not at $\phi = 2\pi$ but at $\phi = 2\pi/\,(1-\epsilon) \simeq 2\pi + 3\pi r^*/\alpha$. Namely, the perihelion advances (i.e. the whole orbit rotates in the same sense as the planet itself) per revolution by (Fig. 6.6)

$$\delta\phi = \frac{3\pi r^*}{\alpha} = \frac{3\pi r^*}{(1+e)\,r_\mathrm{min}}. \tag{6.54}$$

With the solar Schwarzschild radius $r^*_\odot = 2.95\,\mathrm{km}$, Mercury's eccentricity $e = 0.206$, and its perihelion $r_\mathrm{min} = 4.6 \times 10^7$ km we have the numerical value of the advance as

$$\delta\phi = 5 \times 10^{-7} \text{ radian/revolution} \tag{6.55}$$

or, $5 \times 10^{-7} \times 180/\pi \times 60 \times 60 = 0.103''$ (arcsecond) per revolution. In terms of the advance per century,

$$0.103'' \times \frac{100 \text{ years}}{\text{Mercury's period of } 0.241 \text{ years}} = 43'' \text{ per century.} \tag{6.56}$$

This agrees with the observational evidence.

This calculation explaining the perihelion advance of the planet Mercury from first principles, and the correct prediction for the bending of starlight around the sun, were all obtained by Albert Einstein in an intense two week period in November, 1915. Afterwards, he wrote to Arnold Sommerfeld in a, by now, famous letter

> This last month I have lived through the most exciting and the most exacting period of my life; and it would be true to say this, it has been the most fruitful. Writing letters has been out of the question. I realized that up until now my field equations of gravitation have been entirely devoid of foundation. When all my confidence in the old theory vanished, I saw clearly that a satisfactory solution could only be reached by linking it with the Riemann variations. The wonderful thing that happened then was that not only did Newton's theory result from it, as a first approximation, but also the perihelion motion of the Mercury, as a second approximation. For the deviation of light by the sun I obtained twice the former amount.

We have already discussed the doubling of the light deflection angle. The topics of Riemannian curvature tensor and the GR field equation (having the correct Newtonian limit) will be taken up in Part III—Sections 11.3 and 12.2, respectively.

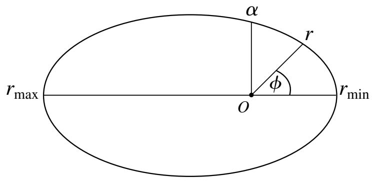

Fig. 6.7 Points on an elliptical orbit are located by the coordinates (r, ϕ), with some notable positions at $(r_{\min}, 0)$, $(r_{\max}, \pi)$, and $(\alpha, \pi/2)$.

Box 6.4 The orbit equation and its perturbation solution

We solve this relativistic energy equation (6.41) as a standard central force problem. The relevant kinematic variables are shown in Fig. 6.7.

The orbit equation. To obtain the orbit equation $r(\phi)$ we first change all the time derivatives into differentiation with respect to the angle ϕ by using the angular momentum equations (6.44) and (6.48):

$$d\tau = \frac{mr^2}{l} d\phi \tag{6.57}$$

and then making the change of variable $u \equiv 1/r$ (and thus $u' \equiv du/d\phi = -u^2(dr/d\phi)$). In this way (6.41) turns into

$$u'^2 + u^2 - \frac{2}{\alpha}u - r^* u^3 = C, \tag{6.58}$$

where $\alpha = l^2/G_N M m^2$, and C is some definite constant. This is the equation we need to solve in order to obtain the planet orbit $r\,(\phi)$.

Zeroth-order solution. Split the solution $u\,(\phi)$ into unperturbed part u_0 and a small correction: $u = u_0 + u_1$ with

$$u_0'^2 + u_0^2 - \frac{2}{\alpha}u_0 = C. \tag{6.59}$$

This unperturbed orbit equation can be solved by differentiating with respect to ϕ and dividing the resultant equation by $2u'$:

$$u_0'' + u_0 = \alpha^{-1}, \tag{6.60}$$

which is a simple harmonic oscillator equation in the variable $(u_0 - \alpha^{-1})$, with ϕ the "time" variable and $\omega = 1$ the "angular frequency." It has the solution $(u_0 - \alpha^{-1}) = D\cos\phi$. We choose to write the constant $D \equiv e/\alpha$ so that the solution takes on the well-known form of a conic section,

$$r = \frac{\alpha}{1 + e\cos\phi}. \tag{6.61}$$

It is clear (see Fig. 6.7) that we have $r = \alpha/(1+e) = r_{\min}$ (perihelion) at $\phi = 0$ and $r = \alpha/(1-e) = r_{\max}$, (aphelion) at $\phi = \pi$. Geometrically, e is called the eccentricity of the orbit. The radial distance at $r(\phi = \pi/2) = \alpha$ can be expressed in terms of perihelion and eccentricity as

$$\alpha = (1+e)r_{\min}. \tag{6.62}$$

Relativistic correction. We now plug $u = u_0 + u_1$ into (6.58)

$$(u_0' + u_1')^2 + (u_0 + u_1)^2 - \frac{2}{\alpha}(u_0 + u_1) - r^*(u_0 + u_1)^3 = C$$

and separate out the leading and the next leading terms (with $u_1 = O(r^*)$):

$$\left(u_0'^2 + u_0^2 - \frac{2}{\alpha}u_0 - C\right) + \left(2u_0'u_1' + 2u_0u_1 - \frac{2}{\alpha}u_1 - r^*u_0^3\right)$$
$$+ O(u_1^2, r^{*2}, u_1 r^*) = 0.$$

After using (6.59), we can then pick out the first-order equation:

$$2u_0'u_1' + 2u_0u_1 - \frac{2}{\alpha}u_1 = r^*u_0^3, \tag{6.63}$$

where

$$u_0 = \frac{1 + e\cos\phi}{\alpha}, \quad u_0' = -\frac{e}{\alpha}\sin\phi. \tag{6.64}$$

The equation for u_1 is then given by

$$-e\sin\phi\frac{du_1}{d\phi} + e\cos\phi\, u_1 = \frac{r^*(1 + e\cos\phi)^3}{2\alpha^2}. \tag{6.65}$$

One can verify that it has the solution

$$u_1 = \frac{r^*}{2\alpha^2}\left[(3 + 2e^2) + \frac{1 + 3e^2}{e}\cos\phi - e^2\cos^2\phi + 3e\,\phi\sin\phi\right]. \tag{6.66}$$

The first two terms have the form of the zeroth-order solution, $(A + B\cos\phi)$; thus they represent unobservably small corrections. The third term, being periodic in ϕ in the same way as the zeroth-order term, is also unimportant. We only need to concentrate on the fourth term which is ever-increasing with ϕ (modulo 2π). Plugging this into $u = u_0 + u_1$, we obtain

$$r = \frac{\alpha}{1 + e\cos\phi + \epsilon e\,\phi\sin\phi}. \tag{6.67}$$

The denominator factor $\epsilon = 3r^*/2\alpha$ being a small quantity, the angular terms in the denominator can be cast in the same form as that for the zeroth order solution (6.61): after approximating $\cos\epsilon\phi \simeq 1$ and $\sin\epsilon\phi \simeq \epsilon\phi$ so that

$$e\cos(\phi - \epsilon\phi) \simeq e(\cos\phi + \epsilon\phi\sin\phi), \tag{6.68}$$

we have the solution (6.67) in a more transparent form

$$r = \frac{\alpha}{1 + e\cos[(1-\epsilon)\phi]} \tag{6.69}$$

as shown in (6.53).

6.4 Black holes

We study here the spacetime structure exterior to any object with its mass so compressed that its radius is smaller than its Schwarzschild radius $r^* = 2G_N M/c^2$. Such objects have been given the evocative name **black holes**, because it is impossible to transmit outwardly any signal, any light, from the region inside the $r = r^*$ surface. This comes about as a black hole gives rise to an infinite gravitational time dilation, that is, a vanishingly small effective speed of light, for an observer far away from the source.

6.4.1 Singularities of the Schwarzschild metric

The Schwarzschild metric,

$$ds^2 = -\left(1 - \frac{r^*}{r}\right) c^2 dt^2 + \left(1 - \frac{r^*}{r}\right)^{-1} dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2), \tag{6.70}$$

has singularities at $r = 0$ and r^*, and[6] $\theta = 0$ and π. We are familiar with the notion that $r = 0$ and $\theta = 0$ and π are **coordinate singularities** associated with our choice of the spherical coordinate system. Namely, they are not physical and do not show up in physical measurements, and they can be transformed away by another coordinate choice. However, in the case here, the $r = 0$ singularity is real. This is not surprising as the Newtonian gravitational potential for a point mass already has this feature: $-G_N M/r$.

What about the $r = r^*$ surface? As we shall demonstrate, it is actually a coordinate singularity, that is, it is not physical and can be transformed away by coordinate transformation (e.g. the Eddington–Finkelstein coordinates, discussed in Box 6.5). However, while physical measurements are not singular at $r = r^*$, it does not mean that this surface is not special. It is an **event horizon**, separating events that can be viewed from afar, from those which cannot (no matter how long one waits). Namely, $r = r^*$ is a boundary of a region, from within it is impossible to send out any signal. It is a boundary of communication, much like earth's horizon is a boundary of our vision.

When studying black hole physics it is often helpful to think of the scenario of a spherical star collapsing into a black hole. The two most important physical effects involved are:

1. The formation of a physical singularity at the end ($r = 0$) of the collapse.
2. As spherical collapse proceeds onward as more of the exterior Schwarzschild geometry is exposed until the formation of the event horizon (at $r = r^*$) restricting communication between an outside observer and the collapsing star, hence preventing the singularity ever being visible.

In the following subsections we shall provide some details for these claims.

6.4.2 Time measurements in the Schwarzschild spacetime

In Chapter 4 we used different inertial coordinate frames (e.g. freely falling spaceship or an observer in the gravitational field watching the spaceship in

[6] The inverse metric has a $[\sin^2\theta]^{-1}$ term.

acceleration, etc.) in order to get different viewpoints of the spacetime. Similarly, useful insights of the Schwarzschild geometry can be had by using different coordinate systems. Here we give the respective descriptions: first according to an observer in a spaceship falling toward the center, then to an observer viewing such event far away from the source.

The local proper time

We already mentioned that $r = r^* \equiv 2G_\mathrm{N}M/c^2$ is a coordinate singularity. Here we will display a specific case of time-measurement by an observer traveling across the Schwarzschild surface. The result shows that such physical measurement is **not** singular at $r = r^*$.

Let τ be the time measured on the surface of a collapsing star (or alternatively, the proper time onboard a spaceship traveling radially toward the origin $r = 0$). Recall from Section 6.3 that, for a particle (with mass) in the Schwarzschild spacetime, we can write a generalized central-force energy equation (6.41). This equation can be simplified further when we specialize to the radial motion of $d\phi/d\tau = 0$ (i.e. $l = 0$) with zero kinetic energy ($\mathcal{K} = 0$) at $r = \infty$ (i.e. the collapsing star or the infalling spaceship start at rest from $r = \infty$):

$$\frac{1}{2}\dot{r}^2 - \frac{G_\mathrm{N}M}{r} = 0 \tag{6.71}$$

or

$$\frac{1}{c^2}\left(\frac{dr}{d\tau}\right)^2 = \frac{2G_\mathrm{N}M}{c^2 r} = \frac{r^*}{r} \quad \text{or} \quad c\,d\tau = \pm\sqrt{\frac{r}{r^*}}dr. \tag{6.72}$$

The + sign corresponds to an exploding star (or, an outward-bound spaceship), while the − sign to a collapsing star (or, an inward-bound probe). So we pick the minus sign. A straightforward integration yields

$$\tau(r) = \tau_0 - \frac{2r^*}{3c}\left[\left(\frac{r}{r^*}\right)^{3/2} - \left(\frac{r_0}{r^*}\right)^{3/2}\right], \tag{6.73}$$

where τ_0 is the time when the probe is at some reference point r_0.

Thus the proper time $\tau(r)$ is perfectly smooth at the Schwarzschild surface, (see Fig. 6.8). The time for the star to collapse from $r = r^*$ to the singular point at $r = 0$ is $\Delta\tau = 2r^*/3c$ which is of the order of 10^{-4} s for a star with a mass 10 times the solar mass (i.e. $r^* \simeq 30\,\mathrm{km}$). An observer on the surface of the collapsing star would not feel anything peculiar when the star passed through the Schwarzschild surface. And it will take both the star and the observer about a tenth of a millisecond to reach the origin, which is a physical singularity.

The Schwarzschild coordinate time

While the time measurement by an observer traveling across the Schwarzschild surface is perfectly finite, this is not the case according to the observer far away from $r = r^*$.

Recall that the coordinate time t in the Schwarzschild spacetime is the time measured by an observer far away from the source, where the spacetime approaches the flat Minkowski space. In the above subsection we have calculated the proper time τ as a function of the radial distance, we now calculate $t(r)$. For this purpose we need to convert (6.72) for the proper time into one for the coordinate time. Using the relation $d\tau = \sqrt{-g_{00}}dt = (1 - r^*/r)^{1/2}dt$

in (6.72) we obtain

$$dt = -\sqrt{\frac{r}{r^*}}\frac{dr}{c(1-(r^*/r))^{1/2}}, \tag{6.74}$$

which can be integrated to yield

$$\begin{aligned} t = t_0 &- \frac{2r^*}{3c}\left[\left(\frac{r}{r^*}\right)^{3/2} - \left(\frac{r_0}{r^*}\right)^{3/2}\right] \\ &+ \frac{r^*}{c}\left\{\ln\left|\frac{\sqrt{(r/r^*)}+1}{\sqrt{(r/r^*)}-1}\cdot\frac{\sqrt{(r_0/r^*)}-1}{\sqrt{(r_0/r^*)}+1}\right| - 2\left[\left(\frac{r}{r^*}\right)^{1/2} - \left(\frac{r_0}{r^*}\right)^{1/2}\right]\right\}. \end{aligned} \tag{6.75}$$

When r and r_0 are much greater than r^*, the coordinate time of (6.75) approaches[7] the proper time of (6.73) as it should. The above logarithmic term can be written as

$$\ln\left|\frac{\sqrt{r}+\sqrt{r^*}}{\sqrt{r}-\sqrt{r^*}}\cdot\frac{\sqrt{r_0}-\sqrt{r^*}}{\sqrt{r_0}+\sqrt{r^*}}\right| = \ln\left|\frac{\left(\sqrt{r}+\sqrt{r^*}\right)^2}{r-r^*}\cdot\frac{r_0-r^*}{(\sqrt{r_0}+\sqrt{r^*})^2}\right|.$$

If r is near r^*, we can drop all nonsingular terms in (6.75) so that

$$t - t_0 = -\frac{r^*}{c}\ln\frac{r-r^*}{r_0-r^*}. \tag{6.76}$$

Equivalently, $(r - r^*) = (r_0 - r^*)e^{-(t-t_0)c/r^*}$. It takes an infinite amount of coordinate time (i.e. the time according to the clock located far from the Schwarzschild surface) to reach $r = r^*$ (see Fig. 6.8).

Infinite gravitational redshift Another way to interpret the above-discussed phenomenon of a distant observer seeing the collapsing star to slow down to a standstill as due to an infinite gravitational time dilation. The relation between coordinate and proper time interval is given by (cf. (6.16)):

$$dt = \frac{d\tau}{\sqrt{-g_{00}}} = \frac{d\tau}{\sqrt{1-(r^*/r)}}. \tag{6.77}$$

The coordinate time interval becomes infinite as r approaches r^*. If we think in terms of wave peaks, it takes an infinite time for the next peak to reach the far away receiver. This can be equivalently phrased as an "infinite gravitational red shift." Our discussion in Section 5.2.2 has

$$\frac{\omega_{\text{rec}}}{\omega_{\text{em}}} = \sqrt{\frac{(g_{00})_{\text{em}}}{(g_{00})_{\text{rec}}}} = \sqrt{\frac{1-(r^*/r_{\text{em}})}{1-(r^*/r_{\text{rec}})}}. \tag{6.78}$$

When $r_{\text{em}} \to r^*$, the received frequency approaches zero, as it would take an infinite interval to receive the next photon (i.e. the peak-to-peak time being proportional to ω^{-1}). Thus no signal transmission from the black hole is possible.

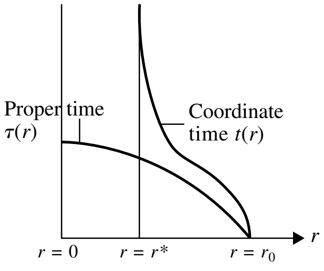

Fig. 6.8 The contrasting behavior of proper time $\tau(r)$ vs. coordinate time $t(r)$ at the Schwarzschild surface.

The "black star" of Michell and Laplace Recall our discussion in Section 3.3.3 that it is tempting to attribute the behavior of a light ray in a gravitational field as due to a "gravitational mass" of the photon. It often leads

[7] A Taylor expansion of the logarithmic factor leads to the cancellation of factors in the $\{\cdots\}$ in (6.75).

to a more familiar derivation of the GR result. For the situation at hand of a photon not being able to reach a far-away observer, one may interpret it as due to the gravitational attraction of the photon by the spherical mass M. The well-known nonrelativistic expression for the escape velocity

$$v_{\text{esc}} = \sqrt{\frac{2G_{\text{N}}M}{r}} = c\sqrt{\frac{r^*}{r}} \tag{6.79}$$

is independent of the object's mass. To the question of how large a ratio of the source mass to radius has to be that even when the object traveling at the speed of light cannot escape, the answer from (6.79) for $v_{\text{esc}} = c$ is just the GR result of $r = r^*$. Historically, discussions of "black star" were carried out along such lines in the eighteenth century by John Michell, as well as by Pierre Laplace. However, we must recognize that, from the perspective of modern gravitational theory (see discussion in Section 3.3.4), this is a conceptually erroneous approach. The impossibility of sending out a light signal from the region inside the Schwarzschild surface is due to infinite gravitational time dilation, rather than a photon having any gravitational mass.

6.4.3 Lightcones of the Schwarzschild black hole

To gain further insight to the event horizon, it is instructive to examine the behavior of the lightcone in the Schwarzschild spacetime. Let us consider a radial ($d\theta = d\phi = 0$) worldline for a photon:

$$ds^2 = -\left(1-\frac{r^*}{r}\right)c^2dt^2 + \left(1-\frac{r^*}{r}\right)^{-1} dr^2 = 0. \tag{6.80}$$

Thus[8]

$$cdt = \pm\frac{dr}{1-(r^*/r)}. \tag{6.81}$$

This can be integrated to obtain, for some reference spacetime point of (r_0, t_0),

$$c(t-t_0) = \pm\left(r - r_0 + r^*\ln\left|\frac{r-r^*}{r_0-r^*}\right|\right), \tag{6.82}$$

or simply,

$$ct = \pm(r + r^*\ln|r-r^*| + \text{ constant}). \tag{6.83}$$

The + and − signs stand for the outgoing and infalling photon world-lines, as shown in Fig. 6.9. To aid our viewing of this spacetime diagram we have drawn in several lightcones in various spacetime regions. We note that for the region far from the source where the spacetime becomes flat, the lightcone approaches the usual form with ±45° sides.

The most prominent feature we notice is that the lightcones "tip over" when crossing the Schwarzschild surface. Instead of opening toward the $t \to \infty$ direction, they tip toward the $r = 0$ line. This can be understood by noting that the roles of space and time are interchanged in Schwarzschild geometry when one moves across the $r = r^*$ surface:

(a) In the spacetime region (I) outside the Schwarzschild surface $r > r^*$, the time and space coordinates have the usual property being timelike $ds_t^2 < 0$ and spacelike $ds_r^2 > 0$ (cf. (6.13) and (6.14)). Namely, the time axis (i.e. perpendicular to the r axis) is timelike, and a fixed-time

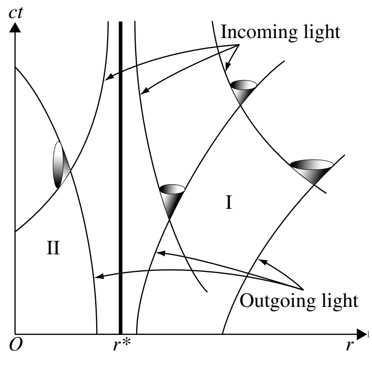

Fig. 6.9 Lightcones in the Schwarzschild spacetime. Regions I and II are separated by the Schwarzschild surface. Different light rays correspond to (6.83) with different constant. Note that the outgoing light ray in region II ends at the $r = 0$ line.

[8] This relation differs from that in (6.74) because we are now considering a lightlike worldline.

worldline is spacelike. Since the trajectory for any real particle must necessarily be timelike[9] and be contained **inside** the lightcone, the lightcones must open toward the $t \to \infty$ direction: an observer in region (I) cannot stop his biological clock.

(b) But in region (II), inside the Schwarzschild surface, their roles are **reversed**. Namely, a worldline of fixed time is now timelike. This comes about because the $(1 - (r^*/r))$ factor changes sign in g_{00} and g_{rr}. For a worldline to remain timelike, an observer can no longer stay put at one position, but is forced to move inward toward the $r = 0$ singularity. For the worldline to be contained within a lightcone, the lightcones themselves must tip over when crossing the $r = r^*$ surface. The tipping over of the lightcones also makes it clear that, once inside the region (II), there is no way one can send a signal to the outside region. Hence, the Schwarzschild surface is an event horizon.

The fact that the metric becomes singular at the $r = r^*$ surface means that the Schwarzschild coordinates, while appropriate for regions far away from the source, are not convenient for the discussion near the Schwarzschild surface. In our description of the "tipping-over" of the lightcones in Fig. 6.9 the use of Schwarzschild coordinates is suspect as the effect is discontinuous across the $r = r^*$ surface. All such doubts are removed when another coordinate system is employed. In the Eddington–Finkelstein coordinates (Box 6.5) the Schwarzschild singularity is removed, and the lightcones, with respect to these new coordinates, tip over smoothly. This also demonstrates explicitly that this is a coordinate singularity, as it is absent in this coordinate system.

Box 6.5 The Eddington–Finkelstein coordinates

The choice of Eddington–Finkelstein coordinates can be motivated as follows. Recall the proper time of an infalling particle into the black hole is smooth for all values of r, cf. (6.73). Thus instead of setting up the coordinate system using a static observer far from the gravitational source (as is the case of the Schwarzschild coordinates) one can describe the Schwarzschild geometry from the viewpoint of an infalling observer. Mathematically, a simpler procedure is to use an infalling photon as the observer to set up the new time coordinate $\bar{t}$. The infalling $ds^2 = 0$ null geodesic in the new $(\bar{t}, r)$ spacetime diagram should be a -45° straight line—just as the infalling photon worldline in the flat spacetime, where the coordinate time is the proper time, cf. (6.14). Such a worldline along radial trajectory $ds^2 = -c^2 d\bar{t}^2 + dr^2 = 0$, or $cd\bar{t} = \pm dr$ is described by the equation

$$c\bar{t} = -r + \text{constant}, \tag{6.84}$$

which should be compared to the equation for an infalling photon in the Schwarzchild coordinates given by (6.83),

$$ct + r^* \ln |r - r^*| = -r + \text{constant}. \tag{6.85}$$

A comparison of the LHSs of (6.84) and (6.85) suggests that we make the coordinate transformation of

$$ct \to c\bar{t} \equiv ct + r^* \ln |r - r^*|. \tag{6.86}$$

[9] The worldline of any real particle remains timelike $ds^2 < 0$ whether outside or inside the Schwarzschild surface. Points along the trajectory must be causally connected (i.e. by signals traveling not faster than light). This feature of lightcone tip over is another way of saying that any light signal emitted inside would follow a worldline that always terminates at $r = 0$.

Differentiating both sides,

$$c d\bar{t} = c dt + \frac{r^*}{r - r^*} dr, \tag{6.87}$$

and substituting into the Schwarzschild line element (6.70) with Ω being the solid angle, we find

$$ds^2 = -\left(1 - \frac{r^*}{r}\right) c^2 d\bar{t}^2 + \frac{r^*}{r} 2c d\bar{t} dr + \left(1 + \frac{r^*}{r}\right) dr^2 + r^2 d\Omega^2, \tag{6.88}$$

which is now regular at $r = r^*$. In fact it is regular in both regions I and II. Thus, this transformation extends the coordinate range[10] from I to both regions I and II. One can object that this extension is achieved by a transformation (6.86) that itself becomes singular at $r = r^*$. However, the only relevant point is that we have found a set of coordinates, as defined by the line element (6.88) which also describes the geometry outside a spherical source. How one found such a set is immaterial.

To look at the lightcone structure in the Eddington–Finkelstein coordinates, we can simplify the algebra by introducing the variable

$$u \equiv c\bar{t} + r. \tag{6.89}$$

Using $du = c d\bar{t} + dr$, the line element in (6.88) can then be written as

$$\begin{aligned} ds^2 &= -c^2 d\bar{t}^2 + dr^2 + \frac{r^*}{r}(c^2 d\bar{t}^2 + 2c d\bar{t} dr + dr^2) + r^2 d\Omega^2 \\ &= -\left(1 - \frac{r^*}{r}\right) du^2 + 2du dr + r^2 d\Omega^2. \end{aligned} \tag{6.90}$$

Thus, for the worldline of a radially ($d\Omega = 0$) infalling photon ($ds^2 = 0$), we must have

$$-\left(1 - \frac{r^*}{r}\right) du^2 + 2du dr = 0. \tag{6.91}$$

Equation (6.91) has two solutions: one being $du = 0$ which is just the straight infalling $-45°$ line of (6.84), forming the left-hand-side edges of the lightcones. The other solution

$$du = \frac{2dr}{1 - (r^*/r)} = c d\bar{t} + dr \tag{6.92}$$

or

$$c\bar{t} = \int \frac{r + r^*}{r - r^*} dr = r + 2r^* \ln|r - r^*| + \text{constant} \tag{6.93}$$

resembles the outward going null line in the Schwarzschild coordinates (6.83), and forms the right-hand-side of the lightcone. Plotting them in Fig. 6.10 we see now that lightcones tip over smoothly across the Schwarzschild surface. Inside the horizon, both sides of lightcones bend toward the $r = 0$ line. In Fig. 6.11, with two spatial dimensions suppressed, we display the spacetime diagram of an imploding star with an observer on its surface sending out light signals at a regular interval.

[10] Another set of coordinates, the Kruskal coordinates, has been discovered; it is valid in even more extensive region than that for the Eddington–Finkelstein coordinates.

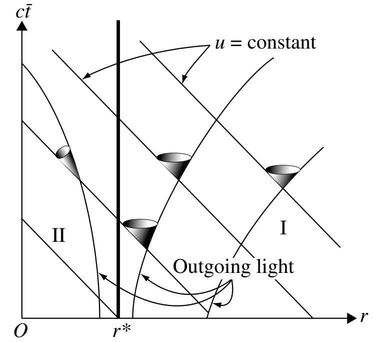

Fig. 6.10 Lightcones in Eddington–Finkelstein spacetime.

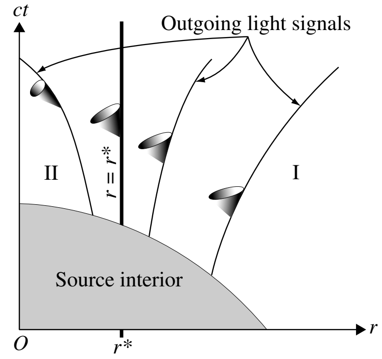

Fig. 6.11 A star undergoing gravitational collapse (two spatial dimensions suppressed). The points on the surface of the collapsing star corresponding to radially moving particles. One such worldline, same as that shown in Fig. 6.8, is displayed above. The region exterior to the collapsing star has the Schwarzschild geometry.

6.4.4 Orbit of an object around a black hole

The formalism presented in our study of the relativistic orbit of a planet can also be applied to the study of the motion of a massive object around a black hole. Let us examine the structure of effective gravitational potential derived in (6.52)

$$\Phi_{\text{eff}} = -\frac{G_{\text{N}}M}{r} + \frac{l^2}{2m^2r^2} - \frac{r^*l^2}{2m^2r^3}. \tag{6.94}$$

While the second term in Φ_{eff} is the familiar centrifugal barrier, the last term is a new GR contribution, which is a small correction for situations such as planet motion, but can be very important when radial distance r is comparable to the Schwarzschild radius r^* as in the case of a compact stellar object. We can find the extrema of this potential by $\partial\Phi_{\text{eff}}/\partial r = 0$

$$\frac{GM}{r^2} - \frac{l^2}{m^2r^3} + \frac{3r^*l^2}{2m^2r^4} = 0 \tag{6.95}$$

or

$$r^2 - \frac{l^2}{GMm^2}r + \frac{3l^2}{2GMm^2}r^* = 0. \tag{6.96}$$

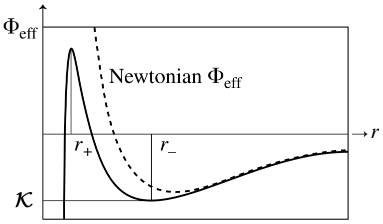

Fig. 6.12 Schwarzschild vs. Newtonian effective potential.

The solutions r_+ and r_- specify the locations where Φ_{eff} has maximum and minimum, respectively, see Fig. 6.12,

$$r_\pm = \frac{l^2}{2GMm^2}\left[1 \mp \left(1 - \frac{6GMm^2}{l^2}r^*\right)\right]^{1/2}. \tag{6.97}$$

We note the distinction from the effective potential in the Newtonian limit of $r^* = 0$: for the Newtonian Φ_{eff} the centrifugal barrier always dominates with $\Phi_{\text{eff}} \to \infty$ in the $r \to 0$ limit, and there is no r_+; a particle cannot fall into the $r = 0$ center as long as $l \neq 0$. In the relativistic Schwarzschild geometry, in the small r limit, the r^* term becomes the most important one and $\Phi_{\text{eff}} \to -\infty$. When $\mathcal{K} \geq m\Phi_{\text{eff}}(r_+)$, a particle can plunge into the gravity center even if $l \neq 0$. If $\mathcal{K} = m\Phi_{\text{eff}}(r_-)$, just like the Newtonian case, we have a stable circular orbit with $r = r_-$. However, this circular radius cannot be arbitrarily small. From (6.97) we have the condition for the circular orbit having the smallest radius:

$$\frac{6GMm^2}{l^2}r^* = 1 \tag{6.98}$$

so that the innermost stable circular orbit has radius

$$r_0 = \frac{l^2}{2GMm^2} = 3r^*. \tag{6.99}$$

6.4.5 Physical reality of black holes

Because of the extraordinary feature of the strongly warped spacetime near the Schwarzschild surface, it took a long time for the physics community to accept the reality of the black hole prediction by the Schwarzschild solution. Here is a short summary of the 50 years development leading to the recognition of the true physical nature of black hole and the modern astronomical observation of such objects.

The long road to the acceptance of black holes' reality

There have been two parallel, and intertwined, lines of study:

1. GR study of the Schwarzschild solution and warped spacetime, much along the lines discussed in our presentation here.
2. Study of gravitational collapse of massive stars—in a normal star, gravitational contraction is balanced by the thermal pressure of the gas, which is large enough if it is hot enough as due to the thermonuclear reactions at the core. The question naturally presents itself: after the exhaustion of nuclear fuel, what will be the fate of a massive star?

We present some of the highlights of this development:

- 1920s and 1930s: No one was willing to accept the extreme predictions that Schwarzschild gave for the highly compact stars. Einstein and Eddington, the opinion setters, openly expressed the view that such gravitational features could not be physical. Calculations were done and results were interpreted as indicating the impossibility of black holes, instead of interpreting them correctly as indicating that no force could resist the gravitational contraction in such a situation.
- 1930: S. Chandrasekhar used the new quantum mechanics to show that, for stellar mass $M > 1.4M_{\odot}$, the electron's **degenerate pressure** will not be strong enough to stop the gravitational contraction. (Electrons obey Pauli's exclusion principle. This effect gives rise to a repulsive force (the degenerate pressure) that resists the gravitational attraction.) Stars having masses under this limit so that the gravitational collapse can be resisted by the electron's fermionic repulsion become **white dwarfs**. In 1932 Chadwick discovered the neutron, which is also a fermion. Zwicky suggested that the remnant of supernova explosion, associated with the final stage of gravitational collapse, was a **neutron star**. Oppenheimer and Volkov, and independently Landau, studied the upper mass limit for neutron stars and found it to be a few solar masses. If greater than this limit, the neutron repulsion would not be large enough to resist the gravitational collapse all the way to the $r = 0$ singularity.
- In the meantime (1939) Oppenheimer and Snyder performed GR study and made most of the points as presented in our discussion here. But the physics community remained skeptical as to the reality of black holes. The reservations were many. For example, one questioned whether the spherical symmetrical situation was too much an idealization? How to take account of the realistic complications such as stellar rotation (the spin causing the star to bulge), deformation to form lumps, shock waves leading to mass ejection, and effects of electromagnetic, gravitational, and neutrino radiation, etc.?
- 1940s and 1950s. The development of atomic and hydrogen bombs during Second World War and the cold war period involved the similar type of physics and mathematical calculations as the study of realistic stellar collapse. From such experience, groups led by Wheeler (USA) and Zel'dovich (USSR) and others carried out realistic simulations. By the end of 1950s, the conclusion had been reached that, despite the complications of spin, deformation, radiation, etc. the implosion proceeded much the way as envisioned in the idealized Oppenheimer and Snyder

calculation. Even with some uncertainty in the nuclear physics involved, this maximum value is determined to be $\approx 2M_\odot$. Any star with a mass $M \gtrsim 2M_\odot$ would contract all the way to become a black hole.

- One development that had a significant impact on the thinking of theorists was the rediscovery in 1958 by Finkelstein of the coordinate system first invented by Eddington (1924) in which the Schwarzschild singularity does not appear, showing clearly that it is a coordinate singularity (see Box 6.5).

Observational evidence of black holes

Black holes being small black discs in the sky far away, it would seem rather hopeless to ever observe them. But by taking account of the gravitational effects of black hole on its surroundings, we now have fairly convincing evidence for a large number of black holes. The basic approach is to determine that the mass of the object is greater than the maximum allowed mass of a neutron star ($\approx 2M_\odot$), then it must be a black hole. The "observed" black holes can be classified into two categories:

1. **Black holes in X-ray binaries**. The majority of all stars are members of binary systems orbiting each other. If the black hole is in a binary system with another visible star, by observing the Kepler motion of the visible companion, one can obtain some limit on the mass of the invisible star as well. If it exceeds $2M_\odot$, it is a black hole candidate. Even better, if the companion star produces significant gas (as is the case of solar flares), the infall of such gas (called **accretion**) into the black hole will produce intense X-rays. A notable example is Cygnus X-1, which is now generally accepted as a black hole binary system with the visible companion having a mass $M_{\text{vis}} > 20M_\odot$ and the invisible black hole having a mass $M > 7M_\odot$. Altogether, close to 10 such binary black holes have been identified in our Galaxy.

2. **Galactic black holes**. It has also been discovered (again by detecting the gravitational influence on visible nearby matter) that at the centers of most galaxies are supermassive black holes, with masses ranging from 10^6 to $10^{12}M_\odot$. Even though the initial finding had been a great surprise, once the discovery was made, it is not too difficult to understand why we should expect such supermassive centers. The gravitational interaction between stars is such that they "swing and fling" past each other: resulting in that one flies off outward while the other falls inward. Thus, we can expect many stars and dust to be driven inward toward the galactic core, producing a supermassive gravitational aggregate. It has been observed that some of these galactic nuclei emit huge amounts of X-rays and visible light to be a thousand times brighter than the stellar light of a galaxy. Such galactic centers are called AGNs (active galactic nuclei). The well-known astrophysical objects, **quasars** (quasi-stellar objects) are interpreted as AGNs in early stage of the cosmic evolution. Observations suggest that an AGN is composed of a massive center surrounded by a molecular accretion disk. They are thought to be powered by rotating supermassive black holes at their cores of such disks. The energy source is ultimately the rest energy of particles. To power such a huge emission one needs extremely efficient mechanisms for releasing the rest energy. Besides the electromagnetic extraction of rotational energy as alluded to above, another important vehicle

is gravitational binding: when free particles falls into lower-energy centrally bound states in the formation of the accretion disk around the black hole. In the following paragraph we briefly discuss the energy release by such a gravitational binding.

Energy release by gravitational binding We are familiar with the fact that thermonuclear fusion is a much more efficient mechanism than chemical reaction to release the (rest) energy mc^2. Here we show that binding of a particle to a compact center of gravity can be an even more efficient mechanism. The thermonuclear reactions taking place in the sun can be summaried as fusing four protons (hydrogen nuclei each with a rest energy of 938 MeV) into a helium nucleus with a released energy of 27 MeV, which represents $27 \div (4 \times 938) \approx 1\%$ of the rest energy. For gravitational binding, consider a free particle, that falls toward a black hole, and ends up bound in a circular orbit (radius r) outside the Schwarzschild radius. The total energy for gravitationally bound particle is given by (Problems 6.4 and 6.5)

$$E = mc^2\left(1 - \frac{r^*}{r}\right)\left(1 - \frac{3}{2}\frac{r^*}{r}\right)^{-1/2}.$$

For the innermost stable circular orbit with $r = 3r^*$ (cf. (6.99)), we have $E = 0.94\ mc^2$. Namely, 6% of the rest energy is released—even larger than thermonuclear fusion.

A glimpse of advanced topics in black hole physics The interested reader is referred to Section A.2 where some advanced topics in black hole physics are very briefly discussed.

Review questions

1. What is the **form** of the spacetime metric (when written in terms of the spherical coordinates) for a spherically symmetric space? Explain very briefly how such a spacetime is curved in space as well as in time.

2. Present a simple proof of **Birkhoff's theorem** for Newtonian gravity. Explain how one then concludes that there is no monopole radiation.

3. Write down the metric function for Schwarzschild spacetime. Given the relation of the metric element g_{00} to the gravitational potential as $-(1 + 2\Phi/c^2)$, demonstrate that the Newtonian result $\Phi = -G_N M/r$ is contained in this solution.

4. How does the feature $g_{rr} = -g_{00}^{-1}$ in the Schwarzschild metric lead to a bending of the light-ray in GR which is twice as much as that predicted by the EP alone, $\delta_{GR} = 2\delta_{EP}$?

5. In simple qualitative terms, explain how gravitational lensing can, in some circumstance, give rise to "Einstein rings," and, in some cases, an enhancement of the brightness of a distant star.

6. Write down the energy equation for the relativistic central force problem used for calculating the precession of the perihelion of the planet Mercury.

7. What does one mean by saying that the Schwarzschild surface is only a coordinate singularity?

8. Explain why the Schwarzschild surface is an "event horizon" (a) by considering gravitational time dilation, and (b) by an examination of the lightcone behavior in the Schwarzschild spacetime (tipping over of the lightcone, etc.).

9. If black holes are invisible, how can we deduce their existence? What are the two classes of black holes for which we already have observational evidence?

Problems

(6.1) **Energy relation for a particle moving in the Schwarzschild spacetime** Show that (6.40), expressing the invariant spacetime interval for a material particle $ds^2 = -c^2 d\tau^2$, can be interpreted as the Schwarzschild spacetime generalization of the familiar special relativistic relation between energy and momentum, $E^2 = p^2c^2 + m^2c^4$ (cf. (2.63)). Namely, show that the flat spacetime ($r^* = 0$) version of (6.40) can be written as $E^2 = p^2c^2 + m^2c^4$.

(6.2) **Equation for a light trajectory** We have used Huygens' principle in Section 3.3.3 and the geodesic equation in Problem 5.2 to derive the expression of gravitational angular deflection $\delta_{\rm GR}$ of (6.29). Here you are asked to obtain this result in yet another way—by following the procedure presented in Section 6.3 when we calculated the orbit equation for a material particle (e.g. the planet Mercury) in the Schwarzschild geometry. Starting with the Lagrangian $L = 0$, instead of $L = -c^2$ (why?), and using the same definition of the two constants of motion as given in (6.44) and (6.45), you have the equation

$$\left(\frac{dr}{d\sigma}\right)^2 + \left(1 - \frac{r^*}{r}\right)\frac{\lambda^2}{4r^2} = c^2\eta^2. \qquad (6.100)$$

Following the same steps as given in Box 6.4, you can change the differentiation with respect to the curve parameter to that of the orbit angle $d\sigma = 2\lambda^{-1}r^2 d\phi$, and use the variable $u = r^{-1}$ to obtain the equation, equivalent to (6.58), for the light trajectory:

$$u'' + u - \epsilon u^2 = 0,$$

where $u'' = d^2u/d\phi^2$ and $\epsilon = 3r^*/2$. A perturbation solution $u = u_0 + \epsilon u_1$ should lead to the result accurate up to the first-order ϵ

$$\frac{1}{r} = \frac{\sin\phi}{r_{\rm min}} + \frac{3 + \cos 2\phi}{4}\frac{r^*}{r_{\rm min}^2}.$$

From this expression for the trajectory $r(\phi)$, one can compare the directions of the initial and final asymptotes to deduce the angular deflection to be $\delta_{\rm GR} = 2r^*/r_{\rm min}$.

(6.3) **Lens equation** Carry out the calculations for (6.36) and (6.37).

(6.4) **Total energy in curved spacetime** Show that the conserved quantity η, as defined by (6.49) $\eta \equiv (1 + 2\mathcal{K}/mc^2)^{1/2}$, has the interpretation of being the total energy per unit rest energy in the Schwarzschild spacetime $\eta = E/mc^2$. Recall that the above quantity $\mathcal{K}$ is the total energy in the nonrelativistic limit, $\mathcal{K} = E - mc^2$.

(6.5) **Circular orbits** For the simplest case of circular orbits, show that the two conserved constants η and l of (6.49) and (6.48) are fixed to be

$$l^2 = G_{\rm N}Mm^2 r\left(1 - \frac{3}{2}\frac{r^*}{r}\right)^{-1},$$

$$\eta^2 = \left(1 - \frac{r^*}{r}\right)^2\left(1 - \frac{3}{2}\frac{r^*}{r}\right)^{-1}. \qquad (6.101)$$

Suggestion Use (6.49) for the bound state total energy $\mathcal{K}/mc^2 = \left(\eta^2 - 1\right)/2$ and write the effective potential (6.52) as

$$\Phi_{\rm eff} = \frac{c^2}{2}\left[\left(1 - \frac{r^*}{r}\right)\left(1 + \frac{l^2}{m^2r^2c^2}\right) - 1\right].$$

(6.6) **Effective speed of light coming out of a black hole vanishes** Following the discussion of gravitational index of refraction in Section 3.3.2 show that, according to an observer far away, the light coming out of a black hole has zero speed.

(6.7) **No stable circular orbit for light around a black hole** Use the effective potential as suggested by the energy balance Eq. (6.100) to show that there is no stable circular trajectory for a photon going around a black hole.

COSMOLOGY

Part II

This page intentionally left blank

7

The homogeneous and isotropic universe

- The framework required to study the whole universe as a physical system is general relativity (GR).
- The universe, when observed on distance scales over 100 Mpc, is homogeneous and isotropic.
- Hubble's discovery that the universe is expanding suggests strongly that it had a beginning when all objects were concentrated at a point of infinite density. The estimate of the age of the universe by astrophysics from observed data is $\gtrsim$12.5 Gyr.
- There is a considerable amount of evidence showing that most of the mass in the universe does not shine. The mass density of the universe, including both luminous and dark matter, is around a third of the "critical density."
- The spacetime satisfying the cosmological principle is described by the Robertson–Walker metric in the comoving coordinates (the cosmic rest frame).
- In an expanding universe with a space that may be curved, any treatment of distance and time must be carried out with care. We study the relations between cosmic redshift and proper, as well as luminosity, distances.

Cosmology is the study of the whole universe as a physical system: what is its matter–energy content? How is this content organized? What is its history? How will it evolve in the future? We are interested in a "smeared" description with the galaxies being the constituent elements of the system. On the cosmic scale the only relevant interaction among galaxies is gravitation; all galaxies are accelerating under their mutual gravity. Thus the study of cosmology depends crucially on our understanding of the gravitational interaction. Consequently, the proper framework for cosmology is GR. The solution of Einstein's equation describes the whole universe because it describes the whole spacetime.

From Chapter 6 we learnt that, for a given gravitational system (M and R being the respectively characteristic mass and length dimensions), one could use the dimensionless parameter

$$\frac{2G_{\mathrm{N}}M}{c^2R} \equiv \psi \tag{7.1}$$

to decide whether Einstein's theory was required, or a Newtonian description would be adequate. In the context of the spatially isotropic solution, it is just the relative size of Schwarzschild radius to the distance scale R. Recall

$\psi_\odot = O(10^{-6})$ for the sun (cf. Eq. (6.22)). Typically the GR effects are small at the level of an ordinary stellar system. On the other hand, we have also considered the case of stellar objects that were so compact that they became black holes when the distance scale is comparable to the Schwarzschild radius $\psi_{\text{bh}} = O(1)$. For the case of cosmology, the mass density is very low. Nevertheless, the distance involved is so large that the total mass M, which increases faster than R, is even larger. This also results in a sizable ψ (Problem 7.1). Thus, to describe events on the cosmic scales, we must use GR concepts.

Soon after the completion of his papers on the foundation of GR, Einstein proceeded to apply his new theory to cosmology. In 1917, he published his paper, "Cosmological considerations on the general theory of relativity." Since then, almost all cosmological studies have been carried out in the framework of GR.

7.1 The cosmos observed

We begin with the observational features of the universe: the organization of its matter content, the large scale motion of its components, its age and mass density.

7.1.1 Matter distribution on the cosmic distance scale

The distance unit traditionally used in astronomy is the parsec (pc). This is defined, see Fig. 7.1(a), as the distance to a star having a parallax of one arcsecond for a base-line equal to the (mean) distance between earth and sun (called an AU, the **astronomical unit**). Thus pc $= (1'' \text{ in radian})^{-1} \times \text{AU} = 3.1 \times 10^{16}\ \text{m} = 3.26$ light-years. (One arcsec equals to 4.85×10^{-6} rad.) Here we first introduce the organization of stars on the cosmic scales of kpc, Mpc, and even hundreds of Mpc.

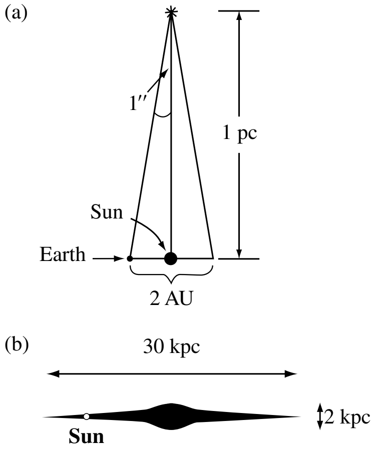

Fig. 7.1 (a) The astronomical distance unit parsec (parallax second) defined, see text. (b) Side view of Milky Way as a typical spiral galaxy.

The distance from the solar system to the nearest star is 1.2 pc. Our own galaxy, the Milky Way, is a typical spiral galaxy. It is comprised of $O(10^{11})$ stars in a disc with a diameter of 30 kpc and a disc thickness of about 2 kpc, see Fig. 7.1(b). Galaxies in turn organize themselves into bodies of increasingly large sizes—into a series of hierarchical clusters. Our galaxy is part of a small cluster, called the Local Group, comprised of about 30 galaxies in a volume measuring 1 Mpc across, for example, the distance to Andromeda galaxy (M31) is 0.7 Mpc. This cluster is part of the Local, or Virgo, Supercluster over a volume measuring 50 Mpc across, with the Virgo cluster comprised of 2000 galaxies over a distance scale of 5 Mpc as its physical center. (The Virgo cluster is about 15 Mpc from us.) This and other clusters of galaxies, such as Hydra–Centaurus supercluster, appear to reside on the edge of great voids. In short, the distribution of galaxies about us is not random, but rather clustered together in coherent patterns that can stretch out up to 100 Mpc. The distribution is characterized by large voids and a network of filamentary structures (see Fig. 7.2). However, beyond this distance scale the universe does appear to be fairly uniform.

7.1.2 Cosmological redshift: Hubble's law

Olbers' paradox: darkness of the night sky Up until less than 100 years ago, the commonly held view was that we lived in a static universe that was infinite in age and infinite in size. However, such a cosmic picture is contradicted by

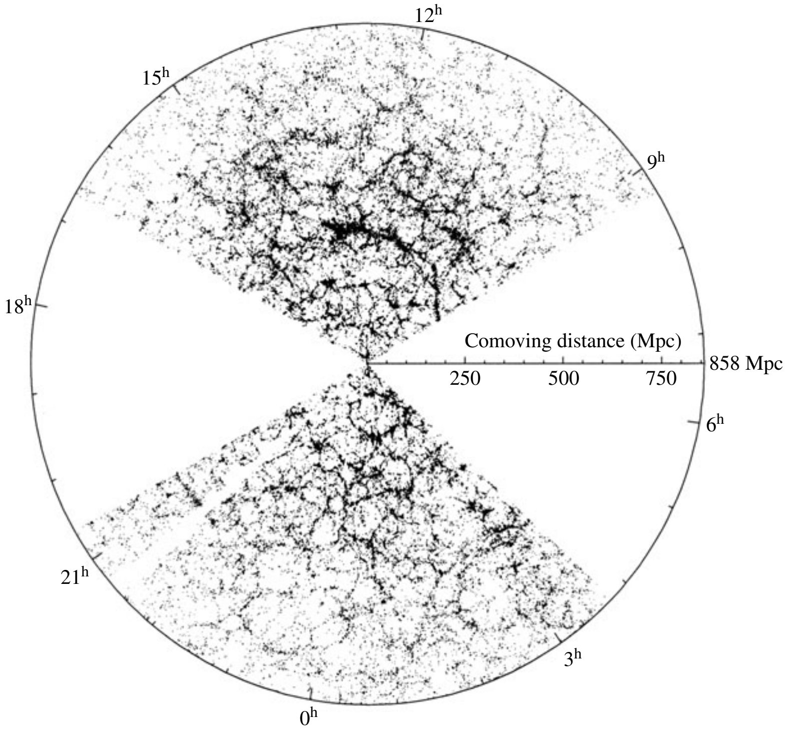

Fig. 7.2 Galaxy distribution out to 858 Mpc, compiled by Gott *et al.* (2003) based on data collected by SDSS and 2dF surveys.

the observation that night sky is dark. If the average luminosity (emitted energy per unit time) of a star is L, then the brightness seen at a distance r would be $f(r) = L/4\pi r^2$. The resultant flux from integrating over all the stars in the infinite universe would be unbounded:

$$B = \int nf(r)dV = nL \int_{r_{\min}}^{\infty} dr = \infty, \tag{7.2}$$

where n, the number density of stars, has been assumed to be a constant. This result of infinite brightness is an over-estimate because stars have finite angular sizes, and the above calculation assumes no obstruction by foreground stars. The correct conclusion is that the night sky in such a universe would have the brightness as if the whole sky were covered by shining suns. Because every line-of-sight has to end at a shining star, although the flux received from a distant star is reduced by a factor of r^{-2} but, for a fixed solid angle, the number of unobstructed stars increases with r^2. Thus, there would be an equal amount of flux from every direction. It is difficult to find any physical mechanism that will allow us to evade this result of night sky ablaze. For example, one might suggest that interstellar dust would diminish the intensity for light having traveled a long distance. But this does not help, because over time, the dust particles would be heated and radiate as much as they absorb.

Maybe our universe is not an infinite and static system?

Hubble's discovery

Astronomers have devised a whole series of techniques that can be used to estimate the distances ever farther into space. Each new one, although less reliable, can be used to reach out further into the universe. During the period of 1910–1930, the "cosmic distance ladder" reached out beyond 100 kpc. The great discovery was made that our universe was composed of a vast collection of galaxies, each resembling our own. One naturally tried to study the motions of these newly discovered "island universes" by using the Doppler effect. When a galaxy is observed at visible wavelengths, its spectrum typically has absorption lines because of the relatively cool upper stellar atmosphere. For a particular absorption line measured in the laboratory to have a wavelength $\lambda_{\rm em}$, the received wavelength by the observer may, however, be different. Such a wavelength shift

$$z \equiv \frac{\lambda_{\rm rec} - \lambda_{\rm em}}{\lambda_{\rm em}} \tag{7.3}$$

is related to the emitter motion by the Doppler effect (cf. Box 10.1), which, for nonrelativistic motion, can be stated as

$$z = \frac{\Delta\lambda}{\lambda} \simeq \frac{v}{c}, \tag{7.4}$$

where v is the recession velocity of the emitter (away from the receiver).

A priori for different galaxies, one expects a random distribution of wavelength shifts: some positive (redshift) and some negative (blueshift). This is more or less true for the Local Group. But beyond the few nearby galaxies, the measurements by Vesto Slipher of some 40 galaxies, over a 10 year period at Arizona's Lowell Observatory, showed that all, except a few in the Local Group, were redshifted. Edwin Hubble (Mt Wilson Observatory, California) then attempted to correlate these redshift results to the more difficult measurements of the distances to these galaxies. The great discovery was made that the redshift was proportional to the distance d to the light emitting galaxy. In 1929, Hubble announced his result:

$$z = \frac{H_0}{c} d \tag{7.5}$$

or, substituting in the Doppler interpretation[1] of (7.4),

$$v = H_0 d. \tag{7.6}$$

Namely, we live in an expanding universe. On distance scales greater than 10 Mpc, all galaxies obey Hubble's law: they are receding from us with speed linearly proportional to the distance. The proportional constant H_0, the **Hubble constant**, gives the recession speed per unit separation (between the receiving and emitting galaxies). To obtain an accurate account of H_0 has been a great challenge as it requires one to ascertain great cosmic distances. Only recently has it become possible to yield consistent results among several independent methods. We have the convergent value[2]

$$H_0 = (72 \pm 7 \text{ km/s})\text{Mpc}^{-1}, \tag{7.7}$$

where the subscript 0 stands for the present epoch $H_0 \equiv H(t_0)$. An inspection of the Hubble's law (7.6) shows that H_0 has the dimension of inverse time, and the measured value in (7.7) can be translated into Hubble time $t_{\rm H} \equiv H_0^{-1} \simeq$ 13.6 Gyr and Hubble length $l_{\rm H} = ct_{\rm H} \simeq$ 4,200 Mpc.

Hubble's law and the Copernican principle

That all galaxies are receding away from us may lead one to suggest erroneously that our location is the center of the universe. The correct

[1] A Doppler redshift comes about because of the increase in the distance between the emitter and the receiver of a light signal. In the familiar situation, this is due to the relative motion of the emitter and the receiver. This language is being used here in our initial discussion of the Hubble's law. However, as we shall show in Sec 7.3, especially Eq. (7.53), the proper description of this enlargement of the cosmic distance as reflecting the expansion of the space itself, rather than the motion of the emitter in a static space.

[2] For a recent compilation of cosmological parameters, see, for example Freedman and Turner (2003).

interpretation is in fact just the opposite. The Hubble relation actually follows naturally from a straightforward extension of the Copernican principle: our galaxy is not at a privileged position in the universe. The key observation is that this is a **linear relation** between distance and velocity at each cosmic epoch. As a result, it is compatible with the same law holding for all observers at every galaxy. Namely, observers on every galaxy would see all the other galaxies receding away from it according to Hubble's law.

Let us write the Hubble's law in a vector form:

$$\mathbf{v} = H_0 \mathbf{r}. \tag{7.8}$$

Namely, a galaxy G, located at position $\mathbf{r}$, will be seen by us (at the origin O) to recede at velocity $\mathbf{v}$ proportional to $\mathbf{r}$. Now consider an observer on another galaxy O′ located at $\mathbf{r}'$ from us as in Fig. 7.3. Then, according to the Hubble's Law, it must be receding from us according to

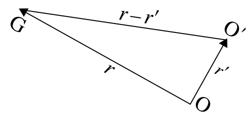

Fig. 7.3 Relative positions of a galaxy G with respect to two observers located at two other galaxies: O and O′.

$$\mathbf{v}' = H_0 \mathbf{r}' \tag{7.9}$$

with the **same** Hubble constant as H_0 is independent of distance and velocity. The difference of these two equations yields

$$(\mathbf{v} - \mathbf{v}') = H_0(\mathbf{r} - \mathbf{r}'). \tag{7.10}$$

But $(\mathbf{r} - \mathbf{r}')$ and $(\mathbf{v} - \mathbf{v}')$ are the respective location and velocity of G as viewed from O′. Since $\mathbf{v}$ and $\mathbf{v}'$ are in the same direction as $\mathbf{r}$ and $\mathbf{r}'$, the vectors $(\mathbf{v} - \mathbf{v}')$ and $(\mathbf{r} - \mathbf{r}')$ must also be parallel. Namely, the relation (7.10) is just the Hubble's law valid for the observer on galaxy O′. Clearly such a deduction would fail if the velocity and distance relation, at a given cosmic time, were nonlinear (i.e. if H_0 depends either on position and/or on velocity).

Distance measurement by redshift

We can turn the Hubble relation around and use it as a means to find the distance to a galaxy by its observed redshift. In fact, the development of new techniques of multi-fiber and multi-slip spectrographs allowed astronomers to measure redshifts for hundreds of galaxies simultaneously. This made large surveys of galaxies possible. In the 1980s there was the Harvard–Smithsonian Center for Astrophysics (CfA) galaxy survey, containing more than 15,000 galaxies. Later, the Las Campanas mapping eventually covered a significantly larger volume and found the "greatness limit" (i.e. cosmic structures have maximum size and on any larger scale the universe would appear to be homogeneous). But this was still not definitive. The modern surveys culminated in two recent parallel surveys: the Anglo-Australian Two-Degree Field Galaxy Redshift Survey (2dF) and the Sloan Digital Sky Survey (SDSS) collaborations have measured some quarter of a million galaxies over a significant portion of the sky, confirming the basic cosmological assumption that the universe of a large distance $\gtrsim$100 Mpc is homogeneous and isotropic. (For further discussion see Sections 7.2 and 7.3.) In fact, an important tool for modern cosmology is just such large-structure study. Detailed analysis of survey data can help us to answer questions such as whether the cosmic structure observed today came about in a top–down (i.e. largest structure formed first, then the smaller ones by fragmentation) or in a bottom–up process. (The second route is favored by observational data.)

In fact many of the cosmological parameters, such as Hubble constant and energy density of the universe, etc. can also be extracted from such analysis.

7.1.3 Age of the universe

If all galaxies are rushing away from each other now, presumably they must have been closer in the past. Unless there was some new physics involved, extrapolating back in time there would be a moment, “the big bang,” when all objects were concentrated at one point of infinite density.[3] This is taken to be the origin of the universe. How much time has evolved since this fiery beginning? What is then the age of our universe?

It is useful to note that the inverse of the Hubble’s constant at the present epoch, the **Hubble time**, has the value of

$$t_{\mathrm{H}} \equiv H_0^{-1} = 13.6 \pm 1.4 \text{ Gyr}. \tag{7.11}$$

By Hubble “constant,” we mean that, at a given cosmic time, H is independent of the separation distance and the recessional velocity—the Hubble relation is a linear relation. The proportional coefficient between distance and recessional speed is not expected to be a constant with respect to time: there is matter and energy in the universe, their mutual gravitational attraction will slow down the expansion, leading to a monotonically decreasing expansion rate $H(t)$—a decelerating universe. Only in an “empty universe” do we expect the expansion rate to be a constant throughout its history, $H(t) = H_0$. In that case, the age t_0 of the empty universe is given by the Hubble’s time

$$[t_0]_{\text{empty}} = \frac{d}{v} = \frac{1}{H_0} = t_{\mathrm{H}}. \tag{7.12}$$

For a decelerating universe full of matter and energy, the expansion rate must be larger in the past: $H(t) > H_0$ for $t < t_0$. Because the universe was expanding faster than the present rate, this would imply that the age of the decelerating universe must be shorter than the empty universe age: $t_0 < t_{\mathrm{H}}$. Nevertheless, we shall often use the Hubble time as a rough benchmark value for the age of the universe, which has a current horizon of $2ct_{\mathrm{H}} = O(10,000 \text{ Mpc})$.

Phenomenologically, we can estimate the age of the universe from observational data. For example, from astrophysical calculation, we know the relative abundance of nuclear elements when they are produced in a star. Since they have different decay rates, their present relative abundance will be different from the initial value. The difference is a function of time. Thus, from the decay rates, the initial and observed relative abundance, we can estimate the time that has elapsed since their formation. Typically, such calculation gives the ages of stars to be around 13 ± 1.5 Gyr. This only gives an estimate of time when stars were first formed, thus only a lower bound for the age of the universe. However, our current understanding informs us that the formation of stars started a hundred million years or so after the big bang, thus such a lower limit still serves an useful estimate of t_0.

An important approach to the study of universe’s age has been the research work on systems of 10^5 or so old stars known as **globular clusters**. These stars are located in the halo, rather than the disc, of our Galaxy. It is known that halo lacks the interstellar gas for star formation. These stars must be created in the early epochs after the big bang (as confirmed by their lack of elements heavier

[3] See Problem 7.9 for a brief description of the alternative cosmology called **steady-state theory** which avoids the big bang beginning by having a constant mass density maintained through continuous spontaneous matter creation as the universe expands.

than lithium, cf. Section 8.4). Stars spend most of their lifetime undergoing nuclear burning. From the observed brightness (flux) and the distance to the stars, one can deduce their intrinsic luminosity (energy output per unit time). From such properties, astrophysical calculations based on established models of stellar evolution, allowed one to deduce their ages (Krauss and Chaboyer, 2003).

$$[t_0]_{\mathrm{gc}} \gtrsim 12.5 \pm 1.5 \text{ Gyr.} \tag{7.13}$$

For reference, we note that the age of our earth is estimated to be around 4.6 Gyr.

7.1.4 Dark matter and mass density of the universe

There is a considerable amount of evidence that most of the mass in the universe does not shine. Namely, in the universe we have **dark matter** as well as luminous. The mass density then has two components:

$$\rho_{\mathrm{M}} = \rho_{\mathrm{LM}} + \rho_{\mathrm{DM}}. \tag{7.14}$$

It is useful to express mass density in terms of a benchmark value for a universe with expansion rate given by the Hubble constant H. One can check that the ratio, with H^2 being divided by the Newton's constant G_{N}, has the units of mass density. With an appropriate choice of coefficient, we have the value of the **critical density**

$$\rho_{\mathrm{c}} = \frac{3H^2}{8\pi G_{\mathrm{N}}}. \tag{7.15}$$

The significance of this quantity will be discussed in Chapter 8 when the Einstein equation for cosmology will be presented. In the meantime, we introduce the notation for the density parameter

$$\Omega \equiv \frac{\rho}{\rho_{\mathrm{c}}}. \tag{7.16}$$

Equation (7.14) for the matter densities may then be written as

$$\Omega_{\mathrm{M}} = \Omega_{\mathrm{LM}} + \Omega_{\mathrm{DM}}, \tag{7.17}$$

where Ω_{LM} and Ω_{DM} are the density parameters for luminous matter and dark matter, respectively. Since the Hubble constant is a function of cosmic time, the critical density also evolves with time. We denote the values for the present epoch with the subscript 0. For example, $\rho(t_0) \equiv \rho_0$, $\rho_{\mathrm{c}}(t_0) \equiv \rho_{\mathrm{c},0}$, and $\Omega(t_0) \equiv \Omega_0$, etc. For the present Hubble constant H_0 as given in (7.7), the critical density has the value

$$\rho_{\mathrm{c},0} = (0.97 \pm 0.08) \times 10^{-29} \text{ g/cm}^3 \tag{7.18}$$

or equivalently a critical energy density[4] of

$$\rho_{\mathrm{c},0}c^2 \simeq 0.88 \times 10^{-10} \text{ J/m}^3 \simeq 5{,}500 \text{ eV/cm}^3. \tag{7.19}$$

In the following, we shall discuss the measurement of the universe's mass density (averaged over volumes on the order of 100 Mpc^3) for both luminous and dark matter. In recent years, these parameters have been deduced rather accurately by somewhat indirect methods: a detailed statistical analysis of the temperature fluctuation in the cosmic microwave background (CMB) radiation and from large structure studies by 2dF and SDSS galaxy surveys mentioned above.

[4] In the natural unit system of quantum field theory, this energy per unit volume is approximately $(2.5 \times 10^{-3}\text{eV})^4/(\hbar c)^3$, where $\hbar$ is Planck's constant (over 2π) with $\hbar c \approx 2 \times 10^{-5}$eV·cm.

The large-structure study involves advanced theoretical tools that are beyond the scope of this introductory presentation. In the following we choose to offer a few methods that involve rather simple physical principles, even though they may be somewhat "dated" in view of recent cosmological advances. Our discussion will, in fact, be only semi-quantitative. Subtle details of derivation, as well as qualification of the stated results, will be omitted. The purpose is to provide some general idea as to how cosmological parameters can in principle be deduced phenomenologically.

Luminous matter

The basic idea of measuring the mass density for the luminous matter is through its relation to the luminosity L (we omit the subscript 0 for the present epoch)

$$\rho_{\rm LM} = \begin{pmatrix}\text{luminosity}\\ \text{density}\end{pmatrix} \times \left(\frac{M}{L}\right). \tag{7.20}$$

Namely, one finds it convenient to decompose mass density into two separate factors: luminosity density and mass-to-luminosity ratio. The luminosity density can be obtained by a count of galaxies per unit volume, multiplied by the average galactic luminosity. Several surveys have resulted in a fairly consistent conclusion of 200 million solar luminosity/Mpc^3,

$$\begin{pmatrix}\text{luminosity}\\ \text{density}\end{pmatrix} \approx 0.2 \times 10^9 \frac{L_\odot}{(\text{Mpc})^3}. \tag{7.21}$$

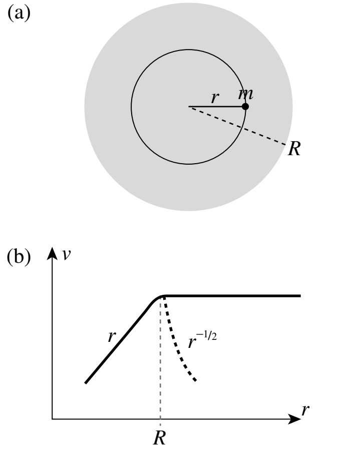

Fig. 7.4 (a) Gravitational attraction on a mass m due to a spherical mass distribution. (b) The velocity $v(r)$ rotation curve (solid line) does not fall as $r^{-1/2}$ beyond R, the edge of the visible portion of a galaxy.

$L_\odot$ is the solar luminosity. The ratio (M/L) is the amount of mass associated, on the average, with a given amount of light. This is the more difficult quantity to ascertain. Depending on the selection criteria one gets a range of values for the mass-to-luminosity ratio. The average of these results came out to be $(M/L) \approx 4M_\odot/L_\odot$. Plugging this and (7.21) into (7.20) we obtain an estimate $\rho_{\rm LM} \approx 8 \times 10^8 M_\odot/\text{Mpc}^3 \approx 5 \times 10^{-32}$ g/cm^3, or a density ratio

$$\Omega_{\rm LM} \approx 0.005. \tag{7.22}$$

Dark matter

Although the dark matter does not emit electromagnetic radiation, it still feels gravitational effects. The most direct evidence of dark matter's existence comes from measured "rotation curves" in galaxies. Consider the gravitational force that a spherical (or ellipsoidal) mass distribution exerts on a mass m located at a distance r from the center of a galaxy, see Fig. 7.4(a). Since the contribution outside the Gaussian sphere (radius r) cancels out, only the interior mass $M(r)$ enters into the Newtonian formula for gravitational attraction. The object is held by this gravity in a circular motion with centripetal acceleration of v^2/r. Hence

$$v(r) = \sqrt{\frac{G_{\rm N} M(r)}{r}}. \tag{7.23}$$

Thus the tangential velocity inside a galaxy is expected to rise linearly with the distance from the center $v \sim r$ if the mass density is approximately constant. For a light source located outside the galactic mass distribution the velocity is expected to decrease as $v \sim 1/\sqrt{r}$, see Fig. 7.4(b). The velocity of particles located at different distances (the rotation curves) can be measured through the 21-cm lines of the hydrogen atoms. However, beyond the visible portion of the galaxies $r > R$, instead of this fall-off, they are observed to stay at the constant peak value as far as the measurement can be made. (See, for example, Cram *et al.*, (1980).) This indicates that the observed object is gravitationally pulled by other than the luminous matter. Such nonluminous matter is believed to form spherical haloes with dimensions considerably larger than the visible disc, see Fig. 7.5. According to (7.23), the flatness of the rotation curve means that $M \propto r$. We can think of the halo as a sphere with mass density decreasing as r^{-2}. Measurements of the rotational curve for spiral galaxies have shown that halo radii are at least ten times larger than the visible radii of the galaxies.

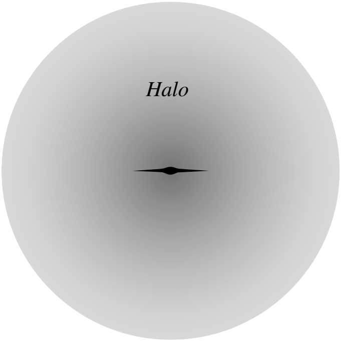

Fig. 7.5 The halo of dark matter surrounding the luminous portion of the galaxy.

Baryonic vs. nonbaryonic matter

- Matter made up of protons and neutrons is generally referred to as "baryonic matter." Baryon is the collective name for strongly interacting particles made up of quark triplets. For our purpose here, the baryon number is just the proton plus neutron numbers. Other types of particles, such as photons, electrons, and neutrinos, carry zero baryon number. Baryon matter can clump to form atoms and molecules, leading to large astronomical bodies. Luminous matter (shining stars) should be baryonic matter; but some of the baryonic matter may not shine—this is the "baryonic dark matter":

$$\Omega_{\mathrm{B}} = \Omega_{\mathrm{LM}} + \Omega_{\mathrm{BDM}}. \tag{7.24}$$

  Nonluminous baryonic matter can be planets or stellar remnants such as black holes, white dwarfs, and brown dwarfs (the last category being stars of the size of Jupiter, with not enough mass to trigger the thermonuclear reaction to make it shine), as well as interstellar gas around galaxies.
- There may also exist gaseous clouds made up of exotic elementary particles that do not have electromagnetic interactions. Neutrinos are cases in point. They only feel the weak nuclear force (i.e. they do not have strong or electromagnetic charges). With their masses being extremely small, neutrinos are expected to be in relativistic motion. They are examples of "**hot dark matter**." There may also be other "weakly interacting massive particles" (WIMPs) that are predicted by various extensions of the standard model of particle interactions.[5] WIMPs, expected to be much more massive than nucleons, are examples of "**cold dark matter**." Hot and cold dark matter have distinctly different effects on the formation of galaxies and clusters of galaxies from initial density inhomogeneity in the universe.[6] Whether hot or cold, such nonbaryonic dark matter will be labeled as "exotic." Exotic particles are necessarily dark.

$$\Omega_{\mathrm{DM}} = \Omega_{\mathrm{BDM}} + \Omega_{\mathrm{exotic}}. \tag{7.25}$$

Namely, the total mass of the universe can be divided into categories, either, depending whether it shines or not, into luminous and dark matter, or depending on their composition, into baryonic and exotic matter:

$$\Omega_{\mathrm{M}} = \Omega_{\mathrm{LM}} + \Omega_{\mathrm{DM}} = \Omega_{\mathrm{B}} + \Omega_{\mathrm{exotic}} \tag{7.26}$$

$$= \Omega_{\mathrm{LM}} + \Omega_{\mathrm{BDM}} + \Omega_{\mathrm{exotic}}. \tag{7.27}$$

[5] It has been suggested that the Standard Model of particle physics be extended by the inclusion of supersymmetry (cf. discussion on p. 282). Every known elementary particle must then have a supersymmetric partner, with a spin differing by half a unit. The lightest of such hypothesized supersymmetry particles are expected to be neutralino ferminos (partners to the neutral Higgs scalar and weak bosons) and should be stable against spontaneous decay. They can in principle make up the bulk of the required dark matter WIMPs.

[6] If the dark matter had been fast moving (hot) particles, they would be able to stream away from high density regions, thus smoothing out small density perturbations. This would have left only the large scale perturbations, leading to the formation of largest structure (superclusters) first, with the smaller structure (galaxies) being produced from fragmentation. However, this top–down scenario is inconsistent with observation.

Thus we need a program to deduce the phenomenological values of Ω_{M} as well as its various components.

The total mass density Ω_{M}

Because the rotation curves cannot be measured far enough out to determine the extent of the dark matter halo, we have to use some other approach to fix the mass density of the dark matter in the universe. Here we discuss one method which allows us to measure the total (luminous and dark) mass in a system of galaxies (binaries, small groups, and large clusters of galaxies), that are bound together by their mutual gravitational attraction. This involves measurements of the mean-square of the galactic velocities $\langle v^2 \rangle$ and the average galactic inverse separation $\langle s^{-1} \rangle$ of, obviously, the luminous components of the system. These two quantities, according to the **virial theorem** of statistical mechanics, $\langle V \rangle = -2\langle T \rangle$, relating the average potential and kinetic energy, are proportional to each other—with the proportional constant given by the total gravitational mass M (luminous and dark) of the system,

$$\langle v^2 \rangle = G_{\mathrm{N}} M \left\langle \frac{1}{s} \right\rangle. \tag{7.28}$$

The proof of this theorem is left as an exercise (Problem 7.6). Here we shall merely illustrate it with a simple example. Consider a two-body system (M, m), with $M \gg m$, separated by distance s. The Newtonian equation of motion $G_{\mathrm{N}} Mm/s^2 = mv^2/s$ immediately yields the result in (7.28). From such considerations, one obtains a total mass density that is something like 80 times larger than the luminous matter. Thus the luminous matter, being what we can see when looking out into space, is only a tiny fraction of the mass content of the universe.

We should add a historical note. That there might be significant amount of dark matter in the universe was first pointed out by Fritz Zwicky in the 1930s. The basis of this proposal is just the method we have outlined here. Zwicky noted that the combined mass of the visible stars and gases in the Coma cluster was simply not enough, given the observed radial velocities of the galaxies, to hold them together gravitationally, that is, what is holding together a galaxy or a cluster of galaxies must be some form of dark matter. The modern era began in 1970 when Vera Rubin and W. Kent Ford, using more sensitive techniques, were able to extend the velocity curve measurements far beyond the visible edge of gravitating systems of galaxies and clusters of galaxies.

There are now several independent means to determine the mass density at the present era $\Omega_{\mathrm{M},0}$: one approach is through gravitational lensing by galaxies, and clusters of galaxies (see Section 6.2), and another is by comparing the number of galaxy clusters in galaxy superclusters throughout the cosmic age. Results, that are generally consistent with the above quoted value have been obtained (Sadoulet, 1999; Griest and Kamionkowski, 2000):

$$\Omega_{\mathrm{M},0} = 0.30 \pm 0.05. \tag{7.29}$$

We shall show in the next chapter that the whole universe is permeated with radiation. However, their energy density is considerably smaller so that $\Omega_{\mathrm{R},0} \ll \Omega_{\mathrm{M},0}$.

The necessity of exotic dark matter

Knowing that mass content of the universe is dominated by dark matter, can we still conclude that most of the matter is baryonic? Namely, can the dark matter, just like the luminous matter, be made up of protons and neutrons? Observational evidence showed that is not the case.

As it turns out, we have methods that can distinguish between baryonic and exotic dark matter because of their different interactions. The light nuclear elements (helium, deuterium, etc.) were produced predominantly in the early universe at the cosmic time $O(10^2\,\mathrm{s})$, cf. Section 8.4. Their abundance (in particular deuterium) is sensitive to the baryonic abundance. From such considerations we have the result (Burles *et al.*, 2001)

$$\Omega_B \simeq 0.04, \tag{7.30}$$

which is confirmed by the latest cosmic microwave anisotropy measurements (see Chapter 9), as well as gravitational microlensing (see Box 6.2). From (7.22), we see that $\Omega_B \gg \Omega_{LM}$. This means that even most of the "ordinary matter" is not visible to us. Our understanding of the baryonic dark matter is still not complete. It is commonly believed that a major portion of it is in the form of unseen ionized gas surrounding galaxies in galactic clusters. Also, with $\Omega_B \ll \Omega_M$ we can conclude that a significant fraction of the dark matter must be exotic:

$$\Omega_{\text{exotic}} = (\Omega_M - \Omega_B) \approx \Omega_M. \tag{7.31}$$

Namely, almost 90% of the matter in the universe is made up of the yet-unknown nonbaryonic dark matter. Most of the speculations have centered around the possibility that such nonbaryonic matter is clouds of weakly interacting massive particles postulated to exist by particle theories that go beyond the standard model verified by current high energy experiments (cf. the discussion leading to (7.25)).

In summary, the total mass density, baryonic and exotic together, is only a third of the critical density:

$$\Omega_M \simeq 0.30 \tag{7.32}$$

most of which is dark

$$\Omega_M = \Omega_{LM} + \Omega_{DM} \qquad \text{with } \Omega_{LM} \approx 0.005. \tag{7.33}$$

Thus, the luminous matter associated with stars and gas we see in galaxies represents about 2% of the total mass content. Most of the matter is dark; the dark matter is in turn composed mostly of exotic particles:

$$\Omega_M = \Omega_B + \Omega_{\text{exotic}} \qquad \text{with } \Omega_B \simeq 0.04. \tag{7.34}$$

The exact nature of these exotic nonbaryonic particles remains one of the unsolved problems in physics.

7.2 The cosmological principle

That the universe is homogeneous and isotropic on the largest scale of hundreds of Mpc has been confirmed by direct observation only very recently (cf. discussion at the end of Section 7.1.2). Another evidence for its homogeneity and isotropy came in the form of extremely uniform CMB radiation. This is the relic thermal radiation left over from an early epoch when the universe was only 10^5 years old. The nonuniformity of CMB is on the order of 10^{-5}. (Cf. Sections 8.5 and 9.3.1.) This shows that the "baby universe" can be described as being highly homogeneous and isotropic.

But long before obtaining this direct observational evidence, Einstein had adopted the **strategy** of starting the study of cosmology with a basic working hypothesis called **the cosmological principle** (CP): at each epoch (i.e. each fixed value of cosmological time t) the universe is homogeneous and isotropic. It presents the **same** aspects (except for local irregularities) from each point.

- This statement that there is no privileged location in the universe (hence homogeneous and isotropic) is sometimes referred to as the **Copernican cosmological principle.**
- This is a priori the most reasonable assumption, as it is difficult to think of any other alternative. Also, in practice, it is also the most "useful," as it involves the least number of parameters. There is some chance for the theory to be predictive. Its correctness can then be checked by observation. Thus CP was invoked in the study of cosmology long before there was any direct observational evidence for a homogeneous and isotropic universe.
- The observed irregularities (i.e. the structure) in the universe-stars, galaxies, clusters of galaxies, superclusters, voids, etc.-are assumed to arise because of gravitational clumping around some initial density unevenness. Various mechanisms for seeding such density perturbation have been explored. Most of the efforts have been concentrated around the idea that, in the earliest moments, the universe passed through a phase of extraordinarily rapid expansion, the "**cosmic inflationary epoch**." The small quantum fluctuations were inflated to astrophysical size and they seeded the cosmological density perturbation (cf. Sections 9.2.3 and 9.3.1).

The cosmological principle gives rise to a picture of the universe as a physical system of "cosmic fluid." The fundamental particles of this fluid are galaxies, and a fluid element has a volume that contains many galaxies, yet is small compared to the whole system of the universe. Thus, the motion of a cosmic fluid element is the smeared-out motion of the constituent galaxies. It is determined by the gravitational interaction of the entire system—the self-gravity of the universe. This means that each element is in free-fall; all elements follow geodesic world-lines. (In reality, the random motions of the galaxies are small, on the order of 10^{-3}.)

Such a picture of the universe allows us to pick a privileged coordinate frame, the **comoving coordinate system**, where

$$t \equiv \text{the proper time of each fluid element}$$

$$x^i \equiv \text{the spatial coordinates carried by each fluid element.}$$

A comoving observer flows with a cosmic fluid element. The comoving coordinate time can be synchronized over the whole system. For example, t is inversely proportional to the temperature of the cosmic background radiation (see Section 8.3) which decreases monotonically. Thus, we can in principle determine the cosmic time by a measurement of the background radiation temperature. This property allows us to define spacelike slices, each with a fixed value of the coordinate time, and each is homogenous and isotropic.

Because each fluid element carries its own position label the comoving coordinate is also the cosmic rest frame—as each fluid element's position coordinates are unchanged with time. But we must remember that in GR the coordinates do not measure distance, which is a combination of the coordinates and the metric. As we shall detail below, the expanding universe, with all galaxies rushing away from each other, viewed in this comoving coordinate, is described not by changing position coordinates, but by an ever-increasing metric. This emphasizes the physics underlying an expanding universe not as something exploding in the space, but as the expansion of space itself.

7.3 The Robertson–Walker metric

The cosmological principle says that, at a fixed cosmic time, each spacelike slice of the spacetime is homogeneous and isotropic. Just as our discussion in Section 6.1 showing that spherical symmetry restricts the metric to the form of $g_{\mu\nu} = \text{diag}(g_{00},\ g_{rr},\ r^2,\ r^2\sin^2\theta)$ with only two scalar functions, g_{00} and g_{rr}, in this section we discuss the geometry resulting from the cosmological principle, which has a Robertson–Walker metric when expressed in the comoving coordinates.

The time components

Because the coordinate time is the proper time of fluid elements, we must have $g_{00} = -1$. The fact that the spacelike slices for fixed t can be defined means that the spatial axes are orthogonal to the time axes:

$$g_{00} = -1 \quad \text{and} \quad g_{0i} = g_{i0} = 0. \tag{7.35}$$

To understand this orthogonality, further details are necessary. Consider an event separated from two other events in distinctive ways: because fixed-time spacelike slices of space exist, we can consider one separation being $da^\mu = (0, dx^i)$ for a definite spatial index i, as well as another separation $db^\mu = (dt, 0)$. The first connects two events on a spacelike space containing all events with the same cosmic time, the second being an interval along the worldline of a comoving observer. The inner product of these two intervals

$$da^\mu db_\mu = g_{i0}dx^i dt \qquad \text{(the } \mu \text{ indices summed, not the } i \text{ indices)}$$

is an invariant, valid in any coordinate system including the local Minkowski frame. This makes it clear that the left-hand side (LHS) vanishes. The above equality then implies $g_{i0} = 0$.

The self-consistency of this choice of coordinates can be checked as follows. A particle at rest in the comoving frame is a particle in free fall under the mutual gravity of the system; it should follow a geodesic worldline obeying (5.9):

$$\frac{d^2x^\mu}{d\tau^2} + \Gamma^\mu_{\alpha\beta}\frac{dx^\alpha}{d\tau}\frac{dx^\beta}{d\tau} = 0. \tag{7.36}$$

Being at rest, $dx^i = 0$ with $i = 1, 2, 3$, we only need to calculate the Christoffel symbol Γ^μ_{00}. But the metric properties of (7.35) imply that

$\Gamma^{\mu}_{00} = 0$. Thus these fluid elements at rest with respect to the comoving frame ($dx^i/d\tau = d^2x^i/d\tau^2 = 0$) do satisfy (trivially) the geodesic equation.

The metric for a 3D space with constant curvature

Let g_{ij} be the spatial part of the metric

$$g_{\mu\nu} = \begin{pmatrix} -1 & 0 \\ 0 & g_{ij} \end{pmatrix} \tag{7.37}$$

that satisfies the cosmological principle. The invariant interval expressed in terms of the comoving coordinates is

$$\begin{aligned} ds^2 &= -c^2dt^2 + g_{ij}dx^idx^j \\ &\equiv -c^2dt^2 + dl^2. \end{aligned} \tag{7.38}$$

Because of the CP requirement (i.e. no preferred direction and position), the time-dependence in g_{ij} must be an **overall** scale factor $R(t)$, with no dependence on any of the indices:

$$dl^2 = R^2(t)d\tilde{l}^2, \tag{7.39}$$

where the reduced length element $d\tilde{l}$ is both t-independent and dimensionless. It is also useful to define a dimensionless scale factor

$$a(t) \equiv \frac{R(t)}{R_0} \tag{7.40}$$

normalized at the present epoch by $a(t_0) = 1$. The denominator on the right-hand side (RHS) $R_0 \equiv R(t_0)$ is sometimes referred to as the radius of the universe now. One has the picture of the universe as a three-dimensional (3D) map with cosmic fluid elements labeled by the fixed comoving coordinates $\hat{x}_i$. Time evolution enters entirely through the time-dependence of the map scale $R(t) = R_0a(t)$, see Fig. 7.6,

$$x_i(t) = R_0a(t)\hat{x}_i \tag{7.41}$$

with $\hat{x}_i$ being the fixed (t-independent) dimensionless map coordinates, while $a(t)$ is the size of the grids and is independent of map coordinates. As the universe expands, the **relative distance** relations (i.e. the shape of things) are not changed.

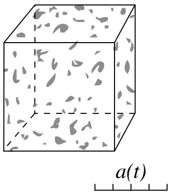

Fig. 7.6 A 3D map of the cosmic fluid with elements labeled by t-independent $\hat{x}_i$ comoving coordinates. The time-dependent of any distance is entirely determined by the t-dependent scale factor $a(t)$.

The Robertson–Walker metric is for a spacetime which, at a give time, has a 3D homogeneous and isotropic space. One naturally expects this 3D space to have a constant curvature. In Section 4.3.2 we have already written down the metric in two spherical coordinate systems:

Equation (4.45) for the comoving coordinates (χ, θ, ϕ):

$$dl^2 = R_0^2a^2(t)\left[d\chi^2 + k^{-1}\left(\sin^2\sqrt{k}\chi\right)d\Omega^2\right]. \tag{7.42}$$

Equation (4.46) for a related comoving spherical system (ξ, θ, ϕ)

$$dl^2 = R_0^2a^2(t)\left(\frac{d\xi^2}{1-k\xi^2} + \xi^2d\Omega^2\right). \tag{7.43}$$

The parameter k in g_{ij} can take on the values $\pm1, 0$ with $k = +1$ for a 3-sphere, $k = -1$ for a 3-pseudosphere, and $k = 0$ for a 3D Euclidean space.

Some of the properties of such spaces, such as their embedding and their volume evaluation, were also discussed in Problems 4.6 and 4.7. In the context of cosmology, the universe having a $k = +1$ positively curved space is called a "**closed universe**," a $k = -1$ negatively curved space an "**open universe**," and $k = 0$ a "**flat universe**." While the deduction of the 3D spatial metric given in Section 4.3.2 is only heuristic, in Section 12.4.1 we shall provide an independent derivation of the same result. In practice, one can use either one of the two coordinates displayed in (7.42) and (7.43); they are equivalent. In the following, for definiteness, we shall work with the (ξ, θ, ϕ) coordinate system of (7.43).

7.3.1 Proper distance in the RW geometry

In an expanding universe with a space that may be curved, we must be very careful in any treatment of distance. In the following sections we shall deal with several kinds of distance, starting with conceptually the simplest: the proper distance.

The proper distance $d_{\rm p}(\xi, t)$ to a point at the comoving radial distance ξ and cosmic time t can be calculated from the metric (7.43) with $d\Omega = 0$ and $dt = 0$.

$$d_{\rm p}(\xi, t) = a(t)R_0 \int_0^{\xi} \frac{d\xi'}{(1 - k\xi'^2)^{1/2}} \tag{7.44}$$

$$= a(t)\left(\frac{R_0}{\sqrt{k}}\right)\sin^{-1}(\sqrt{k}\xi). \tag{7.45}$$

Namely, for a space with positive curvature $k = +1$, we have $a(t)R_0 \sin^{-1}\xi$; negative curvature, $a(t)R_0 \sinh^{-1}\xi$, and a flat space $a(t)R_0\xi$.

This relation

$$d_{\rm p}(\xi, t) = a(t)d_{\rm p}(\xi, t_0) \tag{7.46}$$

implies a proper velocity of

$$v_{\rm p}(t) = \frac{d(d_{\rm p})}{dt} = \frac{\dot{a}(t)}{a(t)}d_{\rm p}(t). \tag{7.47}$$

This is just Hubble's law with the Hubble constant expressed in terms of the scale factor:

$$H(t) = \frac{\dot{a}(t)}{a(t)} \quad \text{and} \quad H_0 = \dot{a}(t_0). \tag{7.48}$$

Recall that the appearance of an overall scale factor in the spatial part of the Robertson–Walker metric follows from our imposition of the homogeneity and isotropy condition. The result in (7.47) confirms our expectation that in any geometrical description of a dynamical universe which satisfies the cosmological principle, Hubble's law emerges automatically. We emphasize that, in the GR framework, the expansion of the universe is described as the expansion of space, and "big bang" is not any sort of "explosion of matter in space," but rather it is an "explosion of space itself." Space is not a "thing" that is expanding, rather space (as represented by the metric function) is the Einstein equation's solution, which has the feature of having an increasing scale factor.

To relate the proper distance to the redshift of a light source located at comoving distance $\xi_{\rm em}$, we use the fact that the observer and emitter are connected by a light ray along a radial path,

$$ds^2 = -c^2dt^2 + R_0^2 a^2(t)\frac{d\xi^2}{1-k\xi^2} = 0.$$

Moving c^2dt^2 to one side and taking the minus sign for the square-root for incoming light, we have

$$R_0 \int_0^{\xi_{\rm em}} \frac{d\xi}{(1-k\xi^2)^{1/2}} = d_{\rm p}(\xi_{\rm em}, t_0) = -\int_{t_0}^{t_{\rm em}} \frac{cdt}{a(t)}, \tag{7.49}$$

where (7.44) has been used to express the first integral in terms of the proper distance at $t = t_0$. The second integral can be put into more useful form by changing the integration variable to the scale factor,

$$-\int_{t_0}^{t_{\rm em}} \frac{cdt}{a(t)} = -\int_1^{a_{\rm em}} \frac{cda}{a(t)\dot{a}(t)} = -\int_1^{a_{\rm em}} \frac{cda}{a^2(t)H(t)}. \tag{7.50}$$

In this way (7.49) becomes the relation between proper distance and scale factor at the emission time

$$d_{\rm p}(\xi_{\rm em}, t_0) = -\int_1^{a_{\rm em}} \frac{cda}{a^2H(a)}. \tag{7.51}$$

7.3.2 Redshift and luminosity distance

We see that the scale factor $a(t)$ is the key quantity in our description of the time evolution of the universe. In fact, because $a(t)$ is generally a monotonic function, it can serve as a kind of cosmic clock. How can the scale factor be measured? The observable quantity that has the simplest relation to $a(t)$ is the wavelength shift of a light signal.

The spectral shift, according to (7.3) is

$$z = \frac{\Delta\lambda}{\lambda} = \frac{\lambda_{\rm rec}}{\lambda_{\rm em}} - 1. \tag{7.52}$$

We expect that the wavelength (in fact any length) scales as $a(t)$ (see Problem 7.8 for a more detailed justification):

$$\frac{\lambda_{\rm rec}}{\lambda_{\rm em}} = \frac{a(t_{\rm rec})}{a(t_{\rm em})}. \tag{7.53}$$

Since the “received time” is at t_0 with $a(t_0) = 1$, we have the basic relation

$$1 + z = \frac{1}{a(t_{\rm em})}. \tag{7.54}$$

For example, at the redshift of $z = 1$, the universe had a size half as large as at the present one. In fact a common practice in cosmology is to refer to “the redshift of an era” instead of its cosmic time. For example, the “photon decoupling time,” when the universe became transparent to light (cf. Section 8.5), is said to occur at $z = 1{,}100$, etc.

Changing the integration variable in (7.51) to the redshift, we have the relation between proper distance and redshift in the Robertson–Walker spacetime:

$$d_{\mathrm{p}}(z) = \int_0^z \frac{c dz'}{H(z')}. \tag{7.55}$$

The functional dependence of distance on the redshift is, of course, the Hubble relation. Different cosmological models having a Hubble constant with different z dependence would yield a different distance-redshift relation. Thus the Hubble curve can be used to distinguish between different cosmological scenarios. As we shall discuss in the next chapter, our universe has been discovered to be in an accelerating expansion phase. By fitting the Hubble curve we shall deduce that the universe's dominant energy component is some unknown "dark energy," which provides the repulsion in causing the expansion to proceed at an ever faster rate.

Luminosity distance and standard candle

The principal approach in calculating the distance to any stellar object is to estimate its true luminosity and compare that with the observed flux (which is reduced by the squared distance). Thus it is important to have stars with known intrinsic luminosity that can be used to gauge astronomical distances. Stars with luminosity that can be deduced from other properties are called "standard candles." A well-known class of standard candles is the Cepheid variable stars, which have a definite correlation between their intrinsic luminosity and their pulse rates. In fact, Edwin Hubble used Cepheids to deduce the distances of the galaxies collected for his distance-vs.-redshift plot. Clearly, the reliability of the method depends on one's ability to obtain the correct estimate of the intrinsic luminosity. A famous piece of history is that Hubble underestimated the luminosity of his Cepheids by almost a factor of 50, leading to an underestimation of the distances, hence an overestimate of the Hubble constant H_0 by a factor of seven. This caused a "cosmic age problem" because the resultant Hubble time (which should be comparable to the age of the universe) became much shorter than the estimated ages of many objects in the universe. This was corrected only after many years of further astronomical observation and astrophysical modeling. Here, we assume that the intrinsic luminosity of a standard candle can be reliably obtained.

In this section, we study the distance that can be obtained by measuring the light flux from a remote light source with known luminosity. Because we use observations of light emitted in the distant past of an evolving universe, this requires us to be attentive in dealing with the concept of time.

The measured flux of watts per unit area is related to the intrinsic luminosity L, which is the total radiated-power by the emitting object, as

$$f = \frac{L}{4\pi d_{\mathrm{L}}^2}. \tag{7.56}$$

This defines the **luminosity distance** d_{L}. In a static universe with a flat geometry, the luminosity distance equals the proper distance to the source: $d_{\mathrm{p(st)}} = d_{\mathrm{L}}$.

$$f = \frac{L}{4\pi d_{\mathrm{p(st)}}^2}. \tag{7.57}$$

In an expanding universe this observed flux, being proportional to energy transfer per unit time, is reduced by a factor of $(1+z)^2$: one power of $(1+z)$ comes from energy reduction due to wavelength lengthening of the emitted light, and another power due to the increasing time interval. Let us explain: The energy being proportional to frequency ω, the emitted energy, compared to the observed one, is given by the ratio,

$$\frac{\omega_{\rm em}}{\omega_0} = \frac{\lambda_0}{\lambda_{\rm em}} = \frac{1}{a(t_{\rm em})} = 1+z, \tag{7.58}$$

where we have used $a(t_0) = 1$ and (7.53) and (7.54). Just as frequency is reduced by $\omega_0 = \omega_{\rm em}(1+z)^{-1}$, the time interval must be correspondingly increased by $\delta t_0 = \delta t_{\rm em}(1+z)$, leading to a reduction of energy transfer rate by another power of $(1+z)$:

$$\frac{\omega_0}{\delta t_0} = \frac{\omega_{\rm em}}{\delta t_{\rm em}}(1+z)^{-2}. \tag{7.59}$$

Thus the observed flux in an expanding universe, in contrast to the static universe result of (7.57), is given by

$$f = \frac{L}{4\pi d_{\rm p}^2 (1+z)^2}. \tag{7.60}$$

Namely, the luminosity distance (7.56) differs from the proper distance by

$$d_{\rm L} = d_{\rm p}(1+z). \tag{7.61}$$

In Chapter 8 the cosmological equations will be solved to obtain the epoch dependent Hubble's constant in terms of the energy/mass content of the universe. In this way we can find how the proper distance $d_{\rm p}$ (thus also the luminosity distance), depends on the redshift z via (7.55) for the general relation. (Problem 7.11 works out the case of small z.) In Box 7.1 we explain the astronomy practice of plotting the Hubble diagrams of redshift vs. **distance modulus** (instead of luminosity distance), which is effectively the logarithmic luminosity distance.

Box 7.1 Logarithmic luminosity and **distance modulus**

Ancient Greek astronomers classified the brightness (observed flux) of stars as having "first magnitude" to "sixth magnitude" for the brightest to the faintest stars visible to the naked eye—the brighter a star is, the smaller its magnitude. Since for this magnitude range of $m_{(6)} - m_{(1)} = 5$ the apparent luminosities span roughly a factor of 100 (namely, $f_{(1)}/f_{(6)} \simeq 100$), a definition of **apparent magnitude** m is suggested:

$$m \equiv -2.5 \log_{10} \frac{f}{f_0}, \tag{7.62}$$

so that $m_{(6)} - m_{(1)} = 2.5 \log_{10}(f_{(1)}/f_{(6)}) = 5$. The reference flux is taken to be $f_0 \equiv 2.52 \times 10^{-8}$ W/m^2 so that the brightest visible stars correspond to $m = 1$ objects. In this scale, for comparison, the sun has an apparent magnitude $m_\odot = -26.8$.

Similar to (7.62), we can define a logarithmic scale, called **absolute magnitude**, for the intrinsic luminosity of a star:

$$M \equiv -2.5 \log_{10} \frac{L}{L_0}, \tag{7.63}$$

where the reference luminosity L_0 is defined so that a star with this power output will be seen at a distance 10 pc away to have a flux f_0:

$$f_0 = \frac{L_0}{4\pi (10\ \mathrm{pc})^2}. \tag{7.64}$$

This works out to be $L_0 = 78.7 L_{\odot}$. Using the definition of luminosity distance as given in (7.56), the Eq. (7.64) can be translated into an expression for the luminosity ratio:

$$\frac{f}{L} = \frac{f_0}{L_0} \left(\frac{10\ \mathrm{pc}}{d_\mathrm{L}} \right)^2. \tag{7.65}$$

Taking the logarithm of this equation leads to the definition of **distance modulus** $(m - M)$, which can be related to luminosity distance by taking the difference of (7.62) and (7.63) and substituting in (7.65):

$$m - M = 5 \log_{10} \frac{d_\mathrm{L}}{10\ \mathrm{pc}}. \tag{7.66}$$

In astronomy literature, one finds the common practice of plotting the Hubble diagram with one axis being the redshift z and another axis, instead of luminosity distance, its logarithmic function, the distance modulus (e.g. Figs 9.8 and 9.11).

Review questions

1. What does it mean that Hubble's law is a linear relation? What is the significance of this linearity? Support your statement with a proof.

2. What is the Hubble time t_H? Under what condition is it equal to the age of the universe t_0? In a universe full of matter and energy, what would be the expected relative magnitude of these two quantities ($t_\mathrm{H} > t_0$ or $t_\mathrm{H} < t_0$)? What is the lower bound for t_0 deduced from the observation data on globular clusters?

3. What are "rotation curves?" What feature would we expect if the luminous matter were a good representation of the total mass distribution? What observational feature of the rotation curve told us that there were significant amounts of dark matter associated with galaxies and clusters of galaxies?

4. Give a simple example that illustrates the content of the virial theorem for a gravitational system. How can this be used to estimate the total mass of the system?

5. What are the values that we have for the total mass density Ω_M, for the luminous matter Ω_LM, and for the baryonic matter Ω_B? From this deduce an estimate of Ω_exotic, the exotic dark matter density parameter. All values are for the present epoch, and list them only to one significant figure.

6. What is the cosmological principle? What are the comoving coordinates?

7. Write out the form of the Robertson–Walker metric for two possible coordinate systems. What is the input (i.e. the assumption) used in the derivation of this metric?

8. What is the physical meaning of the scale factor $a(t)$ and the parameter k in the Robertson–Walker metric? How is the epoch-dependent Hubble constant $H(t)$ related to the scale factor $a(t)$?

9. What is the scaling behavior of wavelength? From this, derive the relation between the scale factor $a(t)$ and the redshift z.

10. Derive the integral expression for the proper distance $d_{\rm p} = c\int H^{-1}dz$ to the light source with redshift z.

11. What is luminosity distance? How is it related to the proper distance?

Problems

(7.1) **The universe as a strong gravitational system** One can check that the universe as a whole corresponds to a system of strong gravity that requires a GR description by making a crude estimate of the parameter ψ in Eq. (7.1). For this calculation you can assume a static Euclidean universe having a finite spherical volume with radius given by a horizon length cH_0^{-1} and having a mass density comparable to the critical density as given in (7.18).

(7.2) **Luminosity distance to the nearest star** The nearest star appears to us to have a brightness $f_* \simeq 10^{-11} f_\odot$ ($f_\odot$ being the observed solar flux). Assuming that it has the same intrinsic luminosity as the sun, estimate the distance d_* to this star, in the distance unit of parsec, as well as in the astronomical unit AU $\simeq 5 \times 10^{-6}$ pc.

(7.3) **Gravitational frequency shift contribution to the Hubble redshift** Hubble's linear plot of redshift vs. distance relies on spectral measurement of galaxies beyond the Local Group with redshift $z \gtrsim 0.01$. A photon emitted by a galaxy suffers not only a redshift because of cosmic recession, but also gravitational redshift. Is the latter a significant factor when compared to the recessional effect? Suggestion: Compare the galactic system with mass $M_G = O(10^{11} M_\odot)$ and size $R_G = O(10^{12} R_\odot)$ to the solar shift of $z_\odot = O(10^{-6})$, *cf*. (3.26).

(7.4) **Energy content due to star light** By assuming the stars have been shining with the same intensity since the beginning of the universe and always had the luminosity density as given in (7.21), estimate the density ratio $\Omega_* = \rho_*/\rho_{\rm c}$ for star light. For this rough calculation you can take the age of universe to be Hubble time $t_{\rm H}$.

(7.5) **Night sky as bright as day** Olbers' paradox is solved in our expanding universe because the age of universe is not infinite $t_0 \simeq t_{\rm H}$ and, having a horizon length $\simeq ct_{\rm H}$, it is effectively not infinite in extent. Given the present luminosity density of (7.21), with the same approximation as Problem 7.4, estimate the total flux due to starlight. Compare your result with the solar flux $f_\odot = L_\odot[4\pi(\text{AU})^2]^{-1}$. We can increase the star light flux by increasing the age of the universe t_0. How much older must the universe be in order that the night sky is a bright as day?

(7.6) **The Virial theorem** Given a general bound system of mass points (located at $\mathbf{r}_n$) subject to gravitational forces (central and inverse square) $\mathbf{F}_n = -\nabla V_n$ with $V_n \propto r_n^{-1}$. By considering the time derivative, and average, of the sum of dot-products of momentum and position $G \equiv \Sigma_n \mathbf{p}_n \cdot \mathbf{r}_n$ (called the **virial**), show that the time-averages of the kinetic and potential energy are related as

$$2\langle T\rangle = -\langle V\rangle.$$

(7.7) **Proper distance from comoving coordinate χ** In the text we worked out the proper distance from a point with radial coordinate ξ as in (7.46). Now perform the same calculation (and obtain a similar result) for a point labeled by the alternative radial coordinate χ with a metric given by (7.42).

(7.8) **Wavelength in an expanding universe** By a careful consideration of the time interval between emission and observation of two successive wavecrests, prove that in an expanding universe with a scale factor $a(t)$ the wavelength scales as expected:

$$\frac{\lambda_{\rm rec}}{\lambda_{\rm em}} = \frac{a(t_{\rm rec})}{a(t_{\rm em})}.$$

Suggestion: cf. Eq. (7.49).

(7.9) **The steady-state universe** In Section 7.3.1 we explained how Hubble's law naturally emerges in any geometric description that satisfies the cosmological principle. The conventional interpretation of an ever increasing scale factor means that all objects must have been closer in the past, leading to a big bang beginning. We also mentioned in Section 7.3.2 that, because of an initial overestimate of the Hubble constant (by a factor of seven), there was a "cosmic age problem." To avoid this difficulty, an alternative cosmology, called the **steady-state universe** (SSU), was proposed by Hermann Bondi, Thomas Gold, and Fred Hoyle. It was suggested that, consistent with the Robertson–Walker description of an expanding universe, cosmological quantities (besides the scale factor)-the expansion rate, deceleration parameter, spatial curvature, matter density, etc.-are all time-independent. A constant mass density means that the universe did not have a big hot beginning; hence there cannot be a cosmic age problem. To have a constant mass density in an expanding universe requires the continuous, energy-nonconserving, creation of matter. To SSUs advocates, this spontaneous mass creation is no more peculiar than the creation of all matter at the instant of big bang. In fact, the name "big bang" was invented by Fred Hoyle as a somewhat disparaging description of the competing cosmology.

(a) Supporters of SSU find this model attractive on theoretical grounds—because it is compatible with the "**perfect cosmological principle**." From the above outline of SSU and the cosmological principle in Section 7.2, can you infer what this "perfect CP" must be?

(b) RW geometry, hence (7.48), also holds for SSU, but with a constant expansion rate $H(t) = H_0$. From this, deduce the explicit t-dependence of the scale factor $a(t)$. What is the SSU prediction for the deceleration parameter q_0 defined in (7.67)?

(c) SSU has a 3D space with a curvature K that is not only constant in space but also in time. Does this extra requirement fix its spatial geometry? If so, what is it?

(d) Since the matter density is a constant $\rho_M(t) = \rho_{M,0} \simeq 0.3\rho_{c,0}$ and yet the scale factor increases with time, SSU requires spontaneous matter creation. What must be the rate of this mass creation per unit volume? Express it in terms of the number of hydrogen atoms created per cubic kilometer per year.

(7.10) **The deceleration parameter and Taylor expansion of the scale factor** Display the Taylor expansion of the scale factor $a(t)$ and $[a(t)]^{-1}$ around $t = t_0$, up to $(t - t_0)^2$, in terms of the Hubble's constant H_0 and the **decleration parameter** defined by

$$q_0 \equiv \frac{-\ddot{a}(t_0)a(t_0)}{\dot{a}^2(t_0)}. \tag{7.67}$$

(7.11) z^2 **correction to the Hubble relation** The Hubble relation (7.5) is valid only in the low velocity limit. Namely, it is the leading term in the power series expansion of the proper distance in terms of the redshift. Using the definition of deceleration parameter introduced in (7.67) one can show that, including the next order, the Hubble relation reads as

$$d_p(t_0) = \frac{cz}{H_0}\left(1 - \frac{1+q_0}{2}z\right). \tag{7.68}$$

(a) One first uses (7.49) to calculate the proper distance up to the quadratic term in the "look-back time" $(t_0 - t_{em})$.

(b) Use the Taylor series of Problem 7.10 to express the redshift in terms of the look-back time up to $(t_0 - t_{em})^2$.

(c) Deduce the claimed result of (7.68) by using the result obtained in (a) and inverting the relation between the redshift and look back time obtained in (b).

8 The expanding universe and thermal relics

- The dynamics of a changing universe are determined by Einstein's equation, which for a Robertson–Walker geometry with the ideal fluid as its mass/energy source takes on the form of Friedmann equations. Through these equations matter/energy determines the scale factor $a(t)$ and the curvature constant k in the metric description of the cosmic spacetime.
- Friedmann equations have simple quasi-Newtonian interpretations.
- The universe began hot and dense (the big bang), and thereafter expanded and cooled. The early universe had undergone a series of thermal equilibria—it had been a set of "cosmic soups" composed of different particles.
- The observed abundance of the light nuclear elements match well with their being the product of the big bang nucleosynthesis.
- When the universe was 350,000 years old photons decoupled, and they remain today as the primordial light having a blackbody spectrum with temperature $T = 2.725$ K.
- The cosmic microwave background (CMB) is not perfectly uniform. The dipole anisotropy is primarily determined by our motion in the rest frame of CMB; higher multipoles contain much information about the geometry, matter/energy content of the universe, as well as the initial density perturbation out of which grew the cosmic structure we see today.

In the previous chapter, we studied the kinematics of the standard model of cosmology. The requirement of a homogeneous and isotropic space fixes the spacetime to have the Robertson–Walker metric in comoving coordinates. This geometry is specified by a curvature signature k and a t-dependent scale factor $a(t)$. Here we study the dynamics of the homogeneous and isotropic universe. The unknown quantities k and $a(t)$ are to be determined by the matter/energy sources through the Einstein field equation as applied to the physical system of the cosmic fluid.

We live in an expanding universe: all the galaxies are now rushing away from each other. This also means that they must have been closer, hence denser and hotter, in the past. Ultimately, at the cosmic beginning $a(0) = 0$, everything must have been right on top of each other. Thus, the standard model of cosmology makes the remarkable prediction that our universe started with a big hot bang.

This prediction that there existed a hotter and denser period in the early universe received strong empirical support, notably by the following discoveries:

1. The 1964 discovery of an all-pervasive microwave background radiation, which is the "after-glow" of the big bang, or the primordial light shining from the early universe, see Section 8.5.
2. The agreement found in the observed abundance of the light nuclear elements, $^4\mathrm{He}, \mathrm{D}, \mathrm{Li}, \ldots$, etc. with the predicted values by the big bang cosmology. The big bang nucleosynthesis will be discussed in Section 8.4.

In Chapter 9, we shall discuss speculations about the nature of the big bang itself, as described by the inflationary cosmology, as well as the recent discovery that the expansion of our universe is accelerating because of the presence of "dark energy," which exerts a repulsive gravitational force.

8.1 Friedmann equations

The Einstein equation relates spacetime's geometry on one side and the mass/energy distribution on another, $G_{\mu\nu} = \kappa T_{\mu\nu}$, cf. Section 5.3.2. For a description of the universe as a physical system that satisfies the cosmological principle, we have learnt in Sections 7.2 and 7.3 that the spacetime must have the Robertson–Walker metric in comoving coordinates. This fixes $G_{\mu\nu}$ on the geometry side of Einstein's equation; the source side should also be compatible with a homogeneous and isotropic space. The simplest plausible choice is to have the energy–momentum tensor $T_{\mu\nu}$ take on the form of an ideal fluid, that is, thermal conductivity and viscosity is unimportant in the cosmic fluid. The proper tensor description of an ideal fluid will be given in Section 10.4 and it is specified by two parameters: mass density ρ and pressure p. Thus the GR field equation relates the geometric parameters of curvature signature k and the scale factor $a(t) = R(t)/R_0$ to the cosmic fluid density $\rho(t)$ and pressure $p(t)$.

The Einstein equation with the Robertson–Walker metric and ideal fluid source leads to the basic set of cosmic equations. They are called the **Friedmann equations**, after the Russian mathematician and meteorologist who was the first, in 1922, to appreciate that the Einstein equation admitted cosmological solutions leading to an expanding universe.[1]

One component of the Einstein equation becomes "the first Friedmann equation,"

$$\frac{\dot{a}^2(t)}{a^2(t)} + \frac{kc^2}{R_0^2 a^2(t)} = \frac{8\pi G_\mathrm{N}}{3}\rho. \tag{8.1}$$

Another component becomes "the second Friedmann equation,"

$$\frac{\ddot{a}(t)}{a(t)} = -\frac{4\pi G_\mathrm{N}}{c^2}\left(p + \frac{1}{3}\rho c^2\right). \tag{8.2}$$

Because the pressure and density factors are positive we have a negative second derivative $\ddot{a}(t)$: the expansion must decelerate because of mutual gravitational attraction among the cosmic fluid elements. It can be shown (Problem 9.1) that

[1] After Einstein's 1917 cosmology paper, one notable contribution was by de Sitter, who studied a dynamical model with nonzero cosmological constant $\Lambda \neq 0$ but devoid of ordinary matter and energy, see Section 9.2.2.

a linear combination of these two Friedmann Eqs (8.1) and (8.2) leads to

$$\frac{d}{dt}(\rho c^2 a^3) = -p\frac{da^3}{dt}, \tag{8.3}$$

which, having the form of the first law of thermodynamics $dE = -pdV$, is the statement of energy conservation. Since it has such a simple physical interpretation, we shall often use Eq. (8.3) instead of (8.2), and by "Friedmann equation" one usually means the first Friedmann Eq. (8.1).

Because there are only two independent equations, yet there are three unknowns functions $a(t), \rho(t)$, and $p(t)$, we need one more relation. This is provided by the "equation of state," relating the pressure to the density of the system. Usually such relation is rather complicated. However, since cosmology deals only with a dilute gas, the equation of state we need to work with can usually be written simply as

$$p = w\rho c^2, \tag{8.4}$$

which defines w as the parameter that characterizes the material content of the system. For example, for nonrelativistic matter the pressure is negligibly small compared to the rest energy of the material particles, hence $w = 0$, and for radiation we have $w = \frac{1}{3}$, etc.

While the precise relation of the Friedmann equations to the Einstein field equation will be discussed in Section 12.4.2, here in Section 8.1.1 we shall present a quasi-Newtonian approach, which gives them a more transparent physical interpretation.

Critical density of the universe

The first Friedmann Eq. (8.1) can be rewritten as

$$-k = \left(\frac{\dot{a}R_0}{c}\right)^2\left(1 - \frac{\rho}{\rho_{\rm c}}\right), \tag{8.5}$$

where the critical density is defined as

$$\rho_{\rm c}(t) = \frac{3}{8\pi G_{\rm N}}\frac{\dot{a}^2}{a^2} = \frac{3[H(t)]^2}{8\pi G_{\rm N}} \tag{8.6}$$

after using the expression of Hubble's constant $H = \dot{a}/a$ of Eq. (7.48). Denoting the density ratio by $\Omega = \rho/\rho_{\rm c}$ as in (8.5), we have

$$-\frac{kc^2}{\dot{a}^2 R_0^2} = 1 - \Omega. \tag{8.7}$$

In particular at $t = t_0$ it becomes, for $\Omega_0 = \rho_0/\rho_{\rm c,0}$,

$$\frac{kc^2}{R_0^2} = H_0^2(\Omega_0 - 1). \tag{8.8}$$

This clearly expresses the GR connection between matter/energy distribution (Ω_0) and geometry (k): if our universe has a mass density greater than the critical density, the average curvature must be positive $k = +1$ (the closed universe);

if the density is less than the critical density, then $k = -1$, the geometry of an open universe having a negative curvature; and if $\rho = \rho_c$, we have the $k = 0$ flat geometry:

$$\begin{array}{llll} \Omega_0 > 1 & \longrightarrow & k = +1 & \text{closed universe,} \\ \Omega_0 = 1 & \longrightarrow & k = 0 & \text{flat universe,} \\ \Omega_0 < 1 & \longrightarrow & k = -1 & \text{open universe.} \end{array} \tag{8.9}$$

Namely, the critical density is the value that separates the positively curved, high-density universe from the negatively curved, low-density universe. From the phenomenological values stated in Chapter 7, $\Omega_0 = \Omega_{M,0} \simeq 0.3$, it would seem that we live in a negatively curved open universe. In Chapter 9 we shall discuss this topic further, and conclude that we need to modify the Einstein equation (by the addition of the cosmological constant). This theoretical input, together with new observational evidence, now suggests that we in fact live in a $k = 0$ flat universe, with the energy/mass density in the universe just equal to the critical value.

Remark: It is often useful to write the Friedmann Eq. (8.1) with the curvature parameter k being replaced by Ω_0 through Eq. (8.8),

$$\frac{H^2}{H_0^2} = \frac{\rho}{\rho_{c,0}} + \frac{1 - \Omega_0}{a^2}. \tag{8.10}$$

8.1.1 The quasi-Newtonian interpretation

The derivation of Friedmann equations involves rather long and tedious calculations (see Section 12.4.2) of the Einstein tensor components $G_{\mu\nu}$ from unknown factors k and $a(t)$ of the metric $g_{\mu\nu}$ and relate them via the Einstein equation $G_{\mu\nu} = \kappa T_{\mu\nu}$ to the density ρ and pressure p from the energy momentum tensor $T_{\mu\nu}$. Having stated this connection of Friedmann Eqs (8.1) and (8.2) to the GR Einstein equation, we now show that they actually have rather simple Newtonian interpretations.

Applicability of Newtonian interpretations

At the beginning of Chapter 7, we presented arguments for the necessity of GR as the proper framework to study cosmology. Indeed, the Friedmann equations are for the scale factor $a(t)$ and the curvature signature k, which are the fundamental concepts of a curved spacetime description of gravity. Nevertheless, as we shall show, these equations have rather simple Newtonian interpretations when supplemented with global geometric concepts at appropriate junctures. There is no contradiction that cosmological equations must fundamentally be relativistic and yet have Newtonian interpretation. The cosmological principle states that every part of the universe, large or small, behaves in the same way. When we concentrate on a small region, Newtonian description should be valid, because gravity involved is not strong and small space can always be approximated by a flat geometry. Thus, we should be able to understand the cosmological equation with a Newtonian approach when it is carried out in an overall GR framework.

Interpretations of the Friedmann equations

Equation (8.3) is clearly the statement of energy conservation as expressed in the form of the first law of thermodynamics $dE = -pdV$. Namely, the change in energy E (with the energy per unit volume ρc^2) is equal to the work done on the system with the volume V being proportional to a^3.

The Friedmann Eq. (8.1) has a straightforward interpretation as the usual energy balance equation (total energy being the sum of kinetic and potential energy) for a central force problem. Recall that in our homogeneous and isotropic cosmological models we ignore any local motions of the galaxies. The only dynamics we need to consider is the change in separation due to the change of the scale factor $a(t)$. Namely, the only relevant dynamical question is the time-dependence of the separation between any two fluid elements.

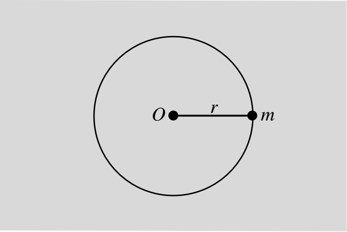

Fig. 8.1 The effect on the separation r between two galaxies due to the gravitational attraction by all the mass in the cosmic fluid. The net force on m is as if all the mass inside the sphere is concentrated at the center O.

To be specific, let us consider a cosmic fluid element (i.e. an element composed of a collection of galaxies), in the homogeneous and isotropic fluid (Fig. 8.1), with mass m at the radial distance $r(\xi, t) = a(t)r_0(\xi)$ with ξ being the dimensionless time-independent comoving radial coordinate, cf. (7.46) with respect to an arbitrarily selected comoving coordinate origin O. We wish to study the effect of gravitational attraction on this mass point m by the whole fluid, which may be treated as spherically symmetric[2] centered around O. The gravitational attraction due to the mass outside the sphere (radius r), in the opposite direction is cancelled. To understand this you can imagine the outside region as composed of a series of concentric spherical shells and the interior gravitational field inside each shell vanishes. This is the familiar Newtonian result. Here we must use GR because the gravitational attraction from the mass shells at large distances is not Newtonian. But it turns out this familiar nonrelativistic solution is also valid in GR, related to the validity of Birkhoff's theorem (as stated at the end of Section 6.1 and Box 12.3). Consequently, the mass element m feels only the total mass M inside the sphere.

This is a particularly simple central force problem, as we have only the radial motion, $\dot{r} = \dot{a}r_0$. The energy balance equation has no orbital angular momentum term:

$$\frac{1}{2}m\dot{r}^2 - \frac{G_N mM}{r} = E_{\text{tot}}, \tag{8.11}$$

which may be re-written as

$$\frac{1}{2}m\dot{a}^2 r_0^2 - \frac{G_N mM}{ar_0} = E_{\text{tot}}. \tag{8.12}$$

The expansion of the universe means an increasing $a(t)$ and potential energy (i.e. it is less negative). This necessarily implies a decreasing $\dot{a}(t)$, namely, a slowdown of the expansion. The total mass inside the (flat space) sphere being $M = \rho 4\pi a^3 r_0^3/3$, we get

$$\dot{a}^2 - \frac{8\pi G_N}{3}\rho a^2 = \frac{2E_{\text{tot}}}{mr_0^2}. \tag{8.13}$$

Remember that this calculation is carried out for an arbitrary center O. Different choices of a center correspond to different values of r_0 and thus different E_{tot}. The assumption of a homogeneous and isotropic space leads to the GR conclusion that the right-hand side (RHS) of Eq. (8.13) is a constant with respect to any choice of O. Furthermore, GR leads to the interpretation of the constant on

[2] Homogeneity and isotropy implies spherical symmetry with respect to every point.

the RHS as the curvature parameter (k):

$$\frac{2E_{\text{tot}}}{mr_0^2} \equiv \frac{-kc^2}{R_0^2}. \tag{8.14}$$

In this way we see that (8.13) is just the first Friedmann Eq. (8.1).

Similarly one can show Friedmann Eq. (8.2) as the $F = ma$ equation of this system (Problem 8.2).

Mass density also determines the fate of the universe

With this interpretation of the Friedmann Eq. (8.1) as the energy balance equation and with the identification of the curvature signature k as being proportional to the total energy of (8.14), it is clear that the value of k, hence also that of density ρ as in (8.5), determines not only the geometry of the 3D space, but also the fate of the cosmic evolution.[3]

For the central force problem, we recall, whether the motion of the test mass m is bound or not is determined by the sign of the total energy E_{tot}. An unbound system allows $r \to \infty$ where the potential energy vanishes and the total energy is given by the kinetic energy, which must be positive: $E_{\text{tot}} > 0$ namely, $k < 0$ (cf. (8.14)). Also the same Eq. (8.11) shows that the sign of E_{tot} reflects the relative size of the positive kinetic energy as compared to the negative potential energy. We can phrase this relative size question in two equivalent ways:

1. **Compare the kinetic energy to a given potential energy: the escape velocity**. Given a source mass (i.e. the potential energy), one can decide whether the kinetic energy term (i.e. test particle velocity) is big enough to have an unbound system. To facilitate this comparison, we write the potential energy term in the form

$$G_{\text{N}}\frac{mM}{r} \equiv \frac{1}{2}mv_{\text{esc}}^2 \tag{8.15}$$

with the **escape velocity** being

$$v_{\text{esc}} = \sqrt{\frac{2G_{\text{N}}M}{r}}. \tag{8.16}$$

The energy Eq. (8.11) then takes the form of

$$G_{\text{N}}\frac{mM}{r}\left(\frac{v^2}{v_{\text{esc}}^2} - 1\right) = E_{\text{tot}}. \tag{8.17}$$

When $v < v_{\text{esc}}$, thus $E_{\text{tot}} < 0$, the test mass m is bound and can never escape.

2. **Compare the potential energy to a given kinetic energy: the critical mass.** Given test-particle's velocity (i.e. the kinetic energy), one can decide whether the potential energy term (i.e. the amount of mass M) is big enough to overcome the kinetic energy to bind the test mass m. Writing the kinetic energy term as

$$\frac{1}{2}m\dot{r}^2 \equiv G_{\text{N}}\frac{mM_{\text{c}}}{r} \tag{8.18}$$

with the **critical mass** being

$$M_{\text{c}} = \frac{r\dot{r}^2}{2G_{\text{N}}} = \frac{a\dot{a}^2r_0^3}{2G_{\text{N}}}, \tag{8.19}$$

[3] Even though this connection between density and the outcome of time evolution is broken when the Einstein equation is modified by the presence of a cosmological constant term as discussed in Chapter 9, the following presentation can still give us some insight to the meaning of the critical density.

the energy Eq. (8.11) then takes the form of

$$\frac{1}{2}m\dot{r}^2\left(1-\frac{M}{M_c}\right)=E_{\text{tot}}. \tag{8.20}$$

When $M > M_c$, thus $E_{\text{tot}} < 0$, the test mass m is bound and can never escape.

The analogous question of whether, given an expansion rate $H(t)$, the test-galaxy m is bound by the gravitational attraction of the cosmic fluid is determined by whether there is enough mass in the arbitrary sphere (on its surface m lies) to prevent m from escaping completely. Namely, the question of whether the universe will expand forever, or its expansion will eventually slow down and re-collapse will be determined by the value of $k \sim -E_{\text{tot}}$. Since the sphere is arbitrary, what matters is the density of the cosmic fluid. We will divide the critical mass (8.19) by the volume of the sphere, and use the Hubble constant relation $H = \dot{a}/a$ of (7.48) to obtain

$$\frac{a\dot{a}^2 r_0^3/2G_{\text{N}}}{4\pi a^3 r_0^3/3}=\frac{3H^2(t)}{8\pi G_{\text{N}}}, \tag{8.21}$$

which is just the critical density ρ_c defined in (8.6). With M/M_c being replaced by ρ/ρ_c, Eq. (8.20) with E_{tot} given by (8.14) is just the Friedmann equation as written in (8.5) and (8.7).

8.2 Time evolution of model universes

We now use Friedmann Eqs (8.1), (8.3) and the equation of state (8.4) to find the time dependence of the scale factor $a(t)$ for a definite value of the curvature k. Although in realistic situations we need to consider several different energy/mass components $\rho = \Sigma_w \rho_w$ with their respective pressure terms $p_w = w\rho_w c^2$, we shall at this stage consider mostly single component systems. To simplify notation we shall omit the subscript w in the density and pressure functions.

Scaling of the density function

Before solving $a(t)$, we shall first study the scaling behavior of density and pressure as dictated by the energy conservation condition (8.3). Carrying out the differentiation in this equation, we have

$$\dot{\rho}c^2=-3(\rho c^2+p)\frac{\dot{a}}{a}, \tag{8.22}$$

which, after using the equation of state (8.4), turns into

$$\frac{\dot{\rho}}{\rho}=-3(1+w)\frac{\dot{a}}{a}. \tag{8.23}$$

This can be solved by straightforward integration:

$$\rho(t)=\rho_0[a(t)]^{-3(1+w)}. \tag{8.24}$$

For a matter dominated universe $w = 0$, and a radiation dominated universe $w = \frac{1}{3}$, the respective densities scale as

$$\rho_{\text{M}}(t)=\rho_{\text{M},0}[a(t)]^{-3} \quad \text{and} \quad \rho_{\text{R}}(t)=\rho_{\text{R},0}[a(t)]^{-4}. \tag{8.25}$$

While the first equation displays the expected scaling behavior of an inverse volume, the second relation can be understood because radiation energy is inversely proportional to wavelength, hence scales as a^{-1}, which is then divided by the volume factor a^3 to get the density. For the special case of negative pressure $p = -\rho c^2$ with $w = -1$, Eq. (8.24) leads to a constant energy density $\rho(t) = \rho(t_0)$ even as the universe expands. As we shall discuss in the next chapter the newly discovered "dark energy" seems to have just this property.

Model universe with $k = 0$

We proceed to solve Eq. (8.1) for the time evolution of the scale factor $a(t)$ for some simple situations. We first consider a class of model universes with $k = 0$. As we shall see, a spatially flat geometry is particularly relevant for the universe we live in. The Friedmann Eq. (8.1) reads

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G_{\text{N}}}{3}\rho_0 a^{-3(1+w)}, \tag{8.26}$$

where we have also plugged in (8.24). Assuming a power law growth for the scale factor

$$a = \left(\frac{t}{t_0}\right)^x \tag{8.27}$$

thus

$$\frac{\dot{a}}{a} = \frac{x}{t}, \tag{8.28}$$

we can immediately relate the age of the universe t_0 to the Hubble time $t_{\text{H}} = H_0^{-1}$:

$$H_0 \equiv \left(\frac{\dot{a}}{a}\right)_{t_0} = \frac{x}{t_0} \quad \text{or} \quad t_0 = x t_{\text{H}}.$$

Also, by substituting (8.28) into (8.26), we see that to match the powers of t on both sides, there must be the relation

$$x = \frac{2}{3(1+w)}. \tag{8.29}$$

For the matter-dominated and radiation-dominated cases

$$\begin{aligned} &\text{MDU } (w = 0) && x = \frac{2}{3} && a = \left(\frac{t}{t_0}\right)^{2/3} && t_0 = \frac{2}{3}t_{\text{H}}, \\ &\text{RDU } \left(w = \frac{1}{3}\right) && x = \frac{1}{2} && a = \left(\frac{t}{t_0}\right)^{1/2} && t_0 = \frac{1}{2}t_{\text{H}}. \end{aligned} \tag{8.30}$$

Also, (8.29) informs us that x is singular for the $w = -1$ universe, which possesses, we have already noted, negative pressure and constant energy density. This means that for this case the scale factor has a nonpower-growth t-dependence, that is, assumption (8.27) is not applicable. We also point out the general situation of $t_0 < t_{\text{H}}$ for $w > -\frac{1}{3}$, and $t_0 > t_{\text{H}}$ for $w < -\frac{1}{3}$.

Here we have considered the specific case of a flat geometry $k = 0$. But we note that the result also correctly describes the early epoch even in a universe with curvature $k \neq 0$. This is so because in the $t \to 0$ limit, the curvature term in the Friedmann Eq. (8.1) is negligible when compared to the $\dot{a}$ term which grows as some negative power of the cosmic time.

Time evolution in single-component universes

Here we shall consider the time evolution of toy universes with only one component of energy/matter. They can be thought as approximations to a more realistic multi-component universe if the energy content is dominated by one component. Namely, a matter-dominated universe means that the energy of the universe resides primarily in the form of (nonrelativistic) matter, even though there may be many more relativistic radiation particles than matter particles.

Radiation-dominated universe The radiation energy density scales as $\rho \sim a^{-4}$, and the Friedmann Eq. (8.1) can be written, with A being some constant, as

$$\dot{a}^2(t) = \frac{A^2}{a^2(t)} - \frac{kc^2}{R_0^2}. \tag{8.31}$$

With a change of variable $y = a^2$ this equation is simplified to

$$\dot{y}^2 + \frac{4kc^2}{R_0^2}y = 4A^2,$$

which has the solutions:

$$a^2(t) = 2At - \frac{kc^2}{R_0^2}t^2. \tag{8.32}$$

We note that this expression for the scale factor does have the correct early universe limit of $a \sim t^{1/2}$ for a radiation-dominated universe, as previously derived in (8.30). The different time dependence of the scale factor $a(t)$ for $k = 0, \pm 1$ is plotted in Fig. 8.2. NB the straight ($\dot{a}$ = constant) line corresponds to an empty universe, and all other curves lie **below** this, reflecting the fact that in all cases the expansion undergoes deceleration because of gravitational attraction.

Matter-dominated universe The matter density scales as $\rho \sim a^{-3}$ and the first Friedmann Eq. (8.1) becomes, with B being also a constant

$$\dot{a}^2(t) = \frac{B}{a(t)} - \frac{kc^2}{R_0^2}. \tag{8.33}$$

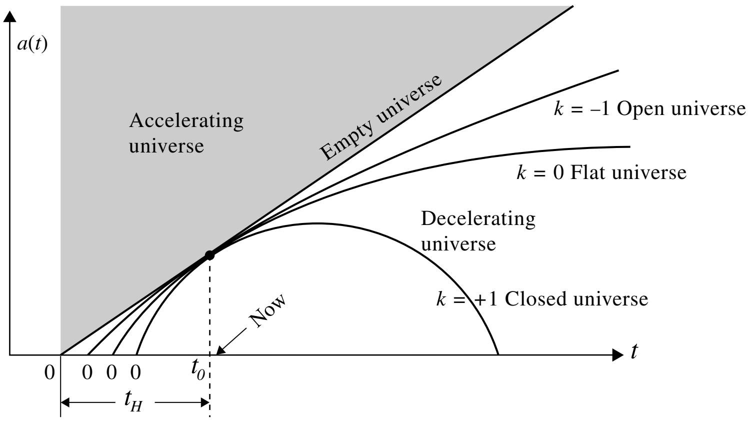

Fig. 8.2 Time dependence of the scale factor $a(t)$ for the open, flat, and closed universe. The qualitative features of these curves are the same for radiation- or matter-dominated universes. All models must have the same radius a_0 and slope $\dot{a}_0$ at t_0 in order to match the Hubble constant $H_0 = \dot{a}_0/a_0$ at the present epoch. The origin of the cosmic time $t = 0$ is different for each curve.

The solution is more complicated. We merely note that the qualitative behavior of $a(t)$ as depicted in Fig. 8.2 is again obtained. Namely, for density less than ρ_c the expansion of the open universe ($k = -1$) will continue forever; for $\rho > \rho_c$ the expansion of a closed universe ($k = +1$) will slow down to a stop then start to recollapse—all the way to another $a = 0$ "big crunch"; for the flat universe ($k = 0$) the expansion will slow down but not enough to stop.

8.3 Big bang cosmology

During the epochs immediately after the big bang, the universe was much more compact, and the energy associated with the random motions of matter and radiation is much larger. Thus, we say, the universe was much hotter. As a result, elementary particles could be in thermal equilibrium through their interactions. As the universe expanded, it also cooled. With the lowering of particle energy, particles (and antiparticles) would either disappear through annihilation, or combine into various composites of particles, or "decouple" to become free particles. As a consequence, there would be different kinds of thermal relics left behind by the hot big bang. One approach to study the universe's history is to start with some initial state which may be guessed based on our knowledge (or speculation) of particle physics. Then we can evolve the universe forward in the hope of ending up with something like the observed universe today. That we can speak of the early universe with any sort of confidence rests with the idea that the universe had been in a series of equilibria (cf. (8.45)). At such stages, the property of the system was determined, independent of the details of the interactions, by a few parameters such as the temperature, density, pressure, etc. Thermodynamical investigation of the cosmic history was pioneered by Tolman (1934). This approach to extract observable consequence from big bang cosmology was first vigorously pursued by George Gamow and his collaborators in the 1940s. Here, we shall give an overview of the thermal history of the universe, in particular, the scale-dependence of radiation temperature.

Once again, it should be pointed out that the calculations carried out in this chapter are rather crude, and they are for illustrative purposes only—to give us a flavor of how in principle cosmological predictions can be made. Typically, realistic calculations would be far more complicated, involving many reaction rates with numerous conditions.

8.3.1 Scale-dependence of radiation temperature

For the radiation component of the universe, we can neglect particle masses and chemical potentials[4] compared to $k_B T$ (here k_B being the Boltzmann's constant). The number distributions with respect to the energy E of the radiation for a gas system composed of bosons (for the minus sign) and fermions (for the positive sign) are, respectively,

$$dn = \frac{4\pi g}{h^3 c^3} \frac{E^2 dE}{e^{E/k_B T} \pm 1}, \tag{8.34}$$

where h is the Planck's constant, and g the number of spin states of the particles making up the radiation. Thus for photons and electrons, we have $g(\gamma) = 2$ and $g(e) = 2$, respectively.[5] Neutrinos have $g(\nu) = 1$ because only the left-handed

[4] Except that for photons, there is no strong theoretical ground to set the chemical potential μ to zero. Yet, since there is nothing that requires a sizable μ, we shall for simplicity set $|\mu| \ll k_B T$ in our discussion.

[5] Particles with mass and spin s have $2s + 1$ spin states (e.g. spin $\frac{1}{2}$ electrons have two spin states), but massless spin particles (e.g. spin 1 photons or spin 2 gravitons) have only two spin states.

states participate in interactions. Carrying out the integration, we get the number density $n = \int dn \sim T^3$:

$$n_b = \frac{4}{3} n_f = 2.404 \frac{g}{2\pi^2} \left(\frac{k_\mathrm{B} T}{hc} \right)^3 \tag{8.35}$$

for the respective boson and fermion systems. We can derive the thermodynamics relation (**Stefan–Boltzmann law**) between radiation energy density and its temperature by performing the integration of $u = \int E dn \sim T^4$:

$$\rho_\mathrm{R} c^2 = u_\mathrm{R} = \frac{g^*}{2} a_\mathrm{SB} T^4, \tag{8.36}$$

where a_SB is the Stefan–Boltzmann constant

$$a_\mathrm{SB} = \frac{\pi^2 k_\mathrm{B}^4}{15 c^3 \hbar^3} = 7.5659 \times 10^{-16}\,\mathrm{J/m^3/K^4}. \tag{8.37}$$

We have summed over the energy contribution by all the constituent radiation particles so that g^* is the "effective number" of spin states of the particles making up the radiation:

$$g^* = \sum_i (g_b)_i + \frac{7}{8} \sum_i (g_f)_i \tag{8.38}$$

with $(g_b)_i$ and $(g_f)_i$ are the spin factors of the ith species of boson or fermion radiation particles, respectively. The $\frac{7}{8}$ factor reflects the different integral values for the fermion distribution, with a plus sign in (8.34), vs. that for the boson case.

Knowing the number and energy densities we can also display the average energy of the constituent radiation particles $\bar{E} = \rho_\mathrm{R} c^2 / n$. In particular we have the photon average energy

$$\bar{E}_\mathrm{R} = 2.7 k_\mathrm{B} T. \tag{8.39}$$

Combining this result (8.36) of $\rho_\mathrm{R} \sim T^4$ with our previous derived relation (8.25) for a radiation-dominated system $\rho_\mathrm{R} \sim a^{-4}$, we deduce the scaling property for the radiation temperature

$$T \propto a^{-1}. \tag{8.40}$$

This expresses, in precise scaling terms, our expectation that temperature is high when the universe is compact, or equivalently, when it expands, it also cools. Under this temperature scaling law, the distributions in (8.34) are unchanged (Tolman, 1934), because the radiation energy was inverse to the wave length, $E \sim \lambda^{-1} \sim a^{-1}(t)$, the combinations $VE^2 dE$ and $E/k_\mathrm{B}T$ were invariant under the scale changes.

Remark: In the context of Newtonian interpretation of the cosmological (Friedmann) equations, we can understand energy conservation in an expanding universe as follows: while the total number of radiation particles $N = nV$ does not change during expansion, the total radiation energy $(k_\mathrm{B} T)$ scales as a^{-1}. This loss of radiation energy, because of an increase in a, is balanced by the increase of gravitational energy of the universe. The gravitational potential energy is also inversely proportional to distance, hence $\sim a^{-1}$, but it is negative. Thus, it increases with an increase in a because it becomes less negative.

Relation between radiation temperature and time The early universe is dominated by radiation with the scale factor $a \propto t^{1/2}$ (cf. (8.30)). We can drop the curvature term in the Friedmann Eq. (8.1), and replace $\dot{a}/a$ by $(2t)^{-1}$ to show that the radiation energy density is related to cosmic time as

$$\rho_{\mathrm{R}} c^2 = \frac{3}{32\pi} \frac{c^2}{G_{\mathrm{N}}} t^{-2}. \tag{8.41}$$

We can rewrite this relation, in a way making it easier to remember, by using the quantum gravity units: Planck energy density and time of (A.9) in Section A.2,

$$\rho_{\mathrm{R}} c^2 = \frac{3}{32\pi} (\rho c^2)_{\mathrm{Pl}} \left(\frac{t_{\mathrm{Pl}}}{t} \right)^2 . \tag{8.42}$$

Namely, in the natural unit system the radiation density is about one tenth per unit cosmic time squared. The radiation density can be related to temperature by the Stefan–Boltzmann law of (8.36), leading to

$$k_{\mathrm{B}} T \simeq 0.46\, E_{\mathrm{Pl}} \left(\frac{t}{t_{\mathrm{Pl}}} \right)^{-1/2} \tag{8.43}$$

or equivalently an easy to remember numerical relation:

$$t(\mathrm{s}) \simeq \frac{10^{20}}{[T(\mathrm{K})]^2}. \tag{8.44}$$

From this estimate we shall see that the big bang nucleosynthesis, taking place at temperature $T_{\mathrm{bbn}} \simeq 10^9$ K (cf. (8.53)), corresponded to a cosmic age of $t_{\mathrm{bbn}} = O(10^2 \text{ s})$ after the big bang.

8.3.2 Different thermal equilibrium stages

Subsequent to the big bang, the cooling of the universe allowed for the existence of different composites of elementary particles. When the falling thermal energy $k_{\mathrm{B}} T$ could no longer produce various types of particle–antiparticle pairs, this lack of fresh supply of antiparticles caused their disappearance from equilibrium states as their annihilation with particles continued. Quarks combined into protons and neutrons (collectively called nucleons). The latter would in turn join into atomic nuclei. At a time some 350,000 years after the big bang, the lower temperature would allow electrons to combine with hydrogen and other light nuclei to form electrically neutral atoms without being immediately blasted apart by high energy electromagnetic radiation. As a result, the universe became transparent to photons. No longer being pushed apart by radiation, the gas of atoms (mostly hydrogen), was free to collapse under its own gravitational attraction, and thus began the process to form stars and galaxies in a background of free photons as we see them today.

In the early universe, energy density was high. This implies a rapidly expanding and cooling universe (cf. (8.1)). To determine what kinds of particle reactions would be taking place to maintain thermal equilibrium involves dynamical calculations, taking into account reaction rate in an expanding and cooling medium. The basic requirement (the “Gamow condition”) is that the reaction rate must be faster than the expansion rate of the system (the universe)

$$\Gamma \geq H, \tag{8.45}$$

where Hubble's constant H measures the expansion rate, and the reaction rate[6] Γ was given by the product of the number density n of the reactant particles, the relative particle velocity v, and the reaction cross section σ, which gives the probability for the reaction to take place:

$$\Gamma = n\sigma v. \tag{8.46}$$

Particle velocity entered because it is the flux of the interacting particles and was given by the velocity distribution as determined by the thermal energy of the system. The condition for a new equilibrium stage to take place is given by the condition of $\Gamma = H$. The cross section being laboratory measured or theory predicted quantities is assumed to be given, and this condition can be used to solve for the thermal energy and the redshift value at which a new equilibrium stage starts.

Epochs of neutrino and positron decoupling

A convenient reference point may be taken when the thermal energy was about 1 GeV corresponding to the age of universe at $t \simeq 10^{-5}$ s. Prior to this, all the stable particles–proton, neutron, electron, neutrino, and their antiparticles, as well as photons–were in thermal equilibrium.

As the universe cooled, different particles would go out of equilibrium. We mention two examples: neutrino decoupling and the disappearance of positrons.

1. The neutrinos started out in thermal equilibrium, through the (reversible) weak interaction reactions, which also allowed proton and neutron to transform into each other.

$$\nu + \mathrm{n} \rightleftarrows \mathrm{e}^- + \mathrm{p}, \quad \bar{\nu} + \mathrm{p} \rightleftarrows \mathrm{e}^+ + \mathrm{n}, \tag{8.47}$$

where n, p, e^-, e^+, ν and $\bar{\nu}$ stand for neutron, proton, electron, positron, neutrino and anti-neutrino, respectively. But as the system cooled, the particle energy was reduced. The cross sections σ of (8.46) for the reactions in (8.47), which had a strong energy-dependence, fell rapidly, and eventually the reaction rate Γ would fall below the expansion rate H. The neutrinos (both neutrinos and anti-neutrinos) no longer interacted with matter. Put in another way, the above listed weak interaction processes, which maintained the neutrinos in thermal equilibrium, and the exchange between protons and neutrons, effectively switched off when the universe cooled below a certain temperature ($k_{\mathrm{B}}T_\nu \approx 3$ MeV). In this way, the neutrinos decoupled from the rest of the matter and the proton–neutron ratio was "frozen".[7] The neutrinos, which participate in weak interactions only, evolved subsequently as free particles. These free neutrinos cooled down as the universe expanded. In the present epoch, the universe should be filled everywhere with these primordial neutrinos (and anti-neutrinos) having a thermal spectrum (with $T_{\nu,0} \approx 1.9$ K) corresponding to a density $n_\nu \approx 150$ cm^{-3}. (See Box 8.1 and similar discussion in Section 8.5 of decoupled photons.) Because the neutrino interaction cross-section is so small, it does not seem possible to detect them with the present technology. Nevertheless, if neutrinos have even a small mass, they can

[6] Properly, the reaction rate should be thermal averaged $\Gamma = n\langle\sigma v\rangle$. Since the product σv is close to a constant, this averaging for our purpose can be done trivially.

[7] Free neutrons can decay into protons (plus electrons and anti-neutrinos) with a half-life just over 10 min. However, soon after nucleon freezeout, before any significant fraction of the neutrons disappeared, they fused into deuterons (each composed of a neutron and proton) and other light nuclei, see Section 8.4. Neutrons bound inside stable nuclei are safe against spontaneous decay if the nuclear binding energy is greater than the neutron–proton rest energy difference.

Table 8.1 A chronology of the universe

		Radiation-dominated universe			Matter-dominated universe		
Cosmic time		10^{-5} s		10^2 s	10^{13} s	10^{15} s	10^{17} s
Thermal energy		1 GeV	3 MeV/1 MeV	0.7 MeV	1 eV		
Age of ...	Quarks-Leptons	Nucleons	ν decouple/ e^+ decouple	Nuclear synthesis	Photon decouple	Galaxies	Now
Physics		Particle physics		Nuclear physics	Atomic physics	Astronomy	

potentially be an important contributor to the (non-baryonic) dark matter mass density of the universe.[8]

2. Similarly, the disappearance of an electron's antiparticles, positrons, proceeded as follows: initially we had the reversible reaction of

$$e^+ + e^- \rightleftarrows \gamma + \gamma. \tag{8.48}$$

However, as the universe cooled, the photons became less energetic. The rest energy of an electron or positron being just over 1/2 MeV, when $k_B T$ fell below their sum of 1 MeV, the reaction could no longer proceed from right to left. Because there were more electrons than positrons, the positrons, by this reaction going from left to right, would be annihilated. They disappeared from the universe. (This matter–antimatter asymmetry, showing up as an excess of electrons over positrons in the universe, will be discussed briefly in Box 8.2 at the end of this chapter.)

When we go back in time, before the age of nucleons, at such high energies the strong interaction underwent a "QCD deconfinement phase transition," when all the quarks inside the nucleons were released.[9] Initially we had mostly the "up" and "down" quarks. As we go further back in time, energy got higher, other heavy flavors of quarks and leptons ("strange" quarks and muons, etc.) would be present, etc. See the **chronology of the universe** (Table 8.1).

In the following sections, we shall discuss two particular epochs in the history of the universe which had left observable features on our present-day cosmos. (1) In Section 8.4 we study the epoch of big bang nucleosynthesis $t_{\text{bbn}} \simeq 10^2$ s, when protons and neutrons combined into charged nuclei (ions) at the end of the age of nucleons. But the lack of stable nuclei with mass number at 5 and 8 prevents the formation of elements heavier than lithium. Thus the abundance of light nuclear elements in the universe can be deduced via cosmological considerations. (2) In Section 8.5 we study the epoch at $t_\gamma \simeq 350{,}000$ year when photons no longer interacted with matter. Having been decoupled, they survived to the present era as the CMB radiation.

8.4 Primordial nucleosynthesis

When we look around our universe, we see mostly hydrogen, and very little of heavy elements. The abundance of heavy elements can all be satisfactorily accounted for by the known nuclear reactions taking place inside stars and supernovae. On the other hand, everywhere we look, besides hydrogen we also

[8] The latest results on neutrino oscillation do indicate that $m_\nu \neq 0$. However, it is so small, $\lesssim 10^{-3}$ eV, that neutrinos cannot possibly be the principal component of dark matter. Furthermore, current understanding of the structure formation in the universe disfavors such "hot dark matter" as the dominant form of dark matter.

[9] The fundamental strong interaction of quantum chromodynamics (QCD), under normal low energy environment, binds quarks into nucleons and other strongly interacting particles. At extremely high energy and density, quarks are set free—deconfined.

see a significant amount of helium. (The helium abundance had been deduced by measurements of the intensities of spectral lines of ^{4}He in stars, planetary nebulas, and H-II regions of galaxies.) The observation data indicate a helium-4 **mass fraction** being close to 24%:

$$y \equiv \left(\frac{^4\mathrm{He}}{\mathrm{H} + {}^4\mathrm{He}} \right)_{\mathrm{mass}} \qquad \text{with } y_{\mathrm{obs}} \simeq 0.24. \tag{8.49}$$

Similarly, we observe a uniform density, at a much smaller abundance, for the light elements: deuterium (D), helium-3 (^{3}He), and lithium-7 (^{7}Li). Gamow and Alpher were the first to suggest, in the late 1940s, that these light nuclear elements were synthesized in the early universe. The primordial processes were theorized to follow the path described below.

The age of nucleons

During this epoch, the cosmic soup was composed of protons, neutrons, electrons, positrons, neutrinos, anti-neutrinos, and photons. There was a tendency for the protons and neutrons to bind (through strong interactions) into nuclear bound states.[10] However, as soon as they were formed, they were blasted apart by energetic photons (photo-dissociation). We can categorize the dominant reactions during the age of nucleons into two types:

1. The transitions between protons and neutrons p ↔ n via prototypical weak interaction processes involving neutrinos as given in (8.47).
2. The protons and neutrons could fuse into light-nuclei ions via strong interaction processes (by adding, one at a time, a proton or a neutron): The key reaction is

$$\mathrm{p} + \mathrm{n} \rightleftarrows \mathrm{D}^+ + \gamma, \tag{8.50}$$

where D$^+$ is the deuteron (i.e., the singly-charged deuterium ion, comprising one proton and one neutron), an isotope of hydrogen, and γ denotes, as before, an energy-carrying photon. As the universe cools there are fewer photons energetic enough to photodissociate the deuteron (the reaction proceeding from the right to the left), and thus deuterons accumulate. The following nucleon capture reactions can then build up heavier elements:

$$\mathrm{D}^+ + \mathrm{n} \rightleftarrows {}^3\mathrm{H}^+ + \gamma, \quad \mathrm{D}^+ + \mathrm{p} \rightleftarrows {}^3\mathrm{He}^{++} + \gamma \tag{8.51}$$

and

$${}^3\mathrm{H}^+ + \mathrm{p} \rightleftarrows {}^4\mathrm{He}^{++} + \gamma, \quad {}^3\mathrm{He}^{++} + \mathrm{n} \rightleftarrows {}^4\mathrm{He}^{++} + \gamma. \tag{8.52}$$

These reversible nuclear reactions would **not** go further, to bind into even heavier nuclei because, helium-4 being a particularly tightly bound nucleus, the formation of a nuclear structure involving 5 nucleons was not energetically favored. Lacking an $A = 5$ stable nucleus, the synthesis of lithium with mass numbers six or seven from stable helium required the much less abundant deuterons or tritium

$${}^4\mathrm{He} + \mathrm{D} \rightleftarrows {}^6\mathrm{Li} + \gamma \quad \text{and} \quad {}^4\mathrm{He} + {}^3\mathrm{H} \rightleftarrows {}^7\mathrm{Li} + \gamma.$$

Big bang nucleosynthesis could not progress further in producing heavier elements ($A > 7$) because there is no stable $A = 8$ element.

[10] A nucleus is composed of Z number of protons and N number of neutrons, giving it the mass number $A = Z + N$. Since chemical properties are determined by the proton number, we can identify Z from the name of the element, for example, hydrogen has $Z = 1$ and helium $Z = 2$. Nuclei having the same Z but different number of neutrons are isotopes. From the mass number, usually denoted by a superscript on the left side of the nucleus symbol, we can figure out the number of neutrons. The most abundant helium isotope is helium-4 (^{4}He) having two neutrons, followed by helium-3 (^{3}He) having one neutron. Hydrogen's isotopes have their distinctive names: the deuteron has one proton and one neutron $^2\mathrm{H} \equiv \mathrm{D}$, and the tritium nucleus ^{3}H has two neutrons.

(If beryllium-8 were formed, it would almost immediately disintegrate into a pair of helium-4s.)

Epoch of primordial nucleosynthesis

But as the universe expanded and cooled, the photons were no longer energetic enough to photodissociate the bound nuclei. This happened, according to a detailed rate calculation, when the thermal energy decreased down to

$$k_B T_{\text{bbn}} \backsimeq 0.7\ \text{MeV}, \tag{8.53}$$

corresponding to an ambient temperature on the order of $T_{\text{bbn}} \backsimeq 10^9$ K, and an age of the universe of $t_{\text{bbn}} \backsimeq O(10^2\ \text{s})$ (cf. (8.41) and (8.44)). The net effect of the above reactions from (8.50) to (8.52) was to cause all the neutrons to be bound into helium-4 nuclei, because there were more protons than neutrons:

$$2\text{n} + 2\text{p} \longrightarrow {}^4\text{He}^{++} + \gamma. \tag{8.54}$$

We can then conclude that the resultant number density n_{He} for helium-4 must be equal to half of the neutron density $n_{\text{He}} = n_{\text{n}}/2$. Use the approximation that helium mass m_{He} is four times the nucleon mass m_{N}, and that the number density of hydrogen n_{H} should equal the proton number density minus neutron density (i.e. they were the left-over protons after all the other protons had combined with the neutrons to form helium ions).

$$\begin{aligned} y \equiv \left(\frac{{}^4\text{He}}{\text{H} + {}^4\text{He}}\right)_{\text{mass}} &= \frac{n_{\text{He}} m_{\text{He}}}{n_{\text{H}} m_{\text{H}} + n_{\text{He}} m_{\text{He}}} \\ &= \frac{(n_{\text{n}}/2) \cdot 4m_{\text{N}}}{(n_{\text{p}} - n_{\text{n}})m_{\text{N}} + (n_{\text{n}}/2) \cdot 4m_{\text{N}}} = \frac{2\lambda}{1+\lambda}, \end{aligned} \tag{8.55}$$

where λ is the neutron to proton ratio: $\lambda = n_{\text{n}}/n_{\text{p}}$. For nonrelativistic nucleons in thermal equilibrium, this ratio is fixed by the Boltzmann distribution, $\exp(-\epsilon/k_B T)$:

$$\lambda = \exp\left[-\left(\frac{\epsilon_{\text{n}} - \epsilon_{\text{p}}}{k_B T_{\text{bbn}}}\right)\right] \backsimeq \exp\left[-\left(\frac{m_{\text{n}} - m_{\text{p}}}{k_B T_{\text{bbn}}}\right)c^2\right] = \text{e}^{-1.3/0.7} = \frac{1}{6.4}, \tag{8.56}$$

where we have used (8.53) and the fact that the neutron is slightly more massive than the proton, with a rest energy difference of $(m_{\text{n}} - m_{\text{p}})c^2 = 1.3$ MeV. After taking into account the fact that some of the neutrons could have decayed into protons via the process $\text{n} \longrightarrow \text{p} + \text{e} + \bar{\nu}$ (but keep in mind that at this stage the free neutron lifetime was still longer than the age of the universe) the neutron to proton ratio λ is reduced somewhat from $\frac{1}{6.4}$ to $\simeq\frac{1}{7}$, yielding a result very close to the observed ratio of 0.24.

$$y = \frac{2\lambda}{1+\lambda} \backsimeq \frac{2/7}{8/7} = \frac{1}{4}. \tag{8.57}$$

In summary, once the deuterium was formed by the fusion of protons and neutrons, this chain of fusion reactions proceeded rapidly so that just about all neutrons were bound into helium. Since these reactions were not perfectly efficient, trace amounts of deuterium and helium-3 were left over. (The small leftover tritium would also decay into helium-3.) Formation of nuclei beyond helium progressed slowly; only small amounts of lithium-6 and -7 were synthesized in the big bang.

Spectral lines of ^{7}Li have been measured in metal-poor stars. Since nuclear elements were produced in stellar thermonuclear reactions, a low abundance of heavy elements indicates that the star was formed from primordial, uncontaminated gas. Thus, at such sites we can assume the elements were produced in big bang nucleosynthesis. Deuterium abundance has been measured in the solar system and in high redshift clouds of interstellar gas by their absorption of light coming from even more distant quasars.

Again, we must keep in mind the crude calculations presented here are for illustrative purpose only. They are meant to give us a simple picture of the physics involved in such cosmological deduction. Realistic calculations often involved the simultaneous inclusion of many reaction rates. In the detailed computations leading to (8.57), one must also use the result of

1. The assumption that there are only three flavors of light neutrinos (electron-, muon-, and tau-neutrinos), because it changes the effective degrees of freedom in the radiation g^* in (8.38);
2. Observed baryon mass density $\rho_{\rm B}$, because it impacts the rate of cooling of the universe. In particular, deuterium D is very sensitive to $\rho_{\rm B}$. Thus, we can use the observed abundance of light elements, in particular deuterium, to determine the baryon density. The best fit, as shown in Fig. 8.3, is at $\rho_{\rm B} \simeq 0.5 \times 10^{-30}$ g/cm^3, or as a fraction of the critical density:

$$\Omega_{\rm B} \simeq 0.044. \tag{8.58}$$

As we already pointed out in Section 7.1.4, when compared to total mass density $\Omega_{\rm M} = \Omega_{\rm B} + \Omega_{\rm exotic} \simeq 0.30$, this shows that baryons are only a small part of the matter in the universe. It also follows that the dark matter must mostly be constituted of unknown exotic weakly interacting massive particles (WIMPs) ($\gg$GeV$/c^2$). Furthermore, we can obtain an estimate of baryon number density $n_{\rm B}$ when we divide the baryon energy density $\rho_{\rm B}c^2 = \Omega_{\rm B} \cdot \rho_{\rm c}c^2 = 220$ MeV/m^3 by the energy of each nucleon, which can be taken to be the rest energy of the nucleon (939 MeV) because these particles are nonrelativistic:

$$n_{\rm B} \simeq 0.23/{\rm m}^3. \tag{8.59}$$

8.5 Photon decoupling and the CMB

The early universe was a plasma of radiation and matter held together by their mutual interaction. As the universe expanded and cooled, matter had congealed into neutral atoms and the cosmic soup lost its ability to entrap the photons. These free thermal photons survived as the CMB radiation we see today. The uniformly distributed relic photons obey a blackbody spectrum. Their discovery gave strong support to the hot big bang beginning of our universe, as it is difficult to think of any other alternative account for the existence of such physical phenomena on the cosmic scale. Furthermore, its slight temperature fluctuation, the CMB anisotropy, is a picture of the "baby universe." Careful study of this anisotropy has furnished and will continue to provide us with

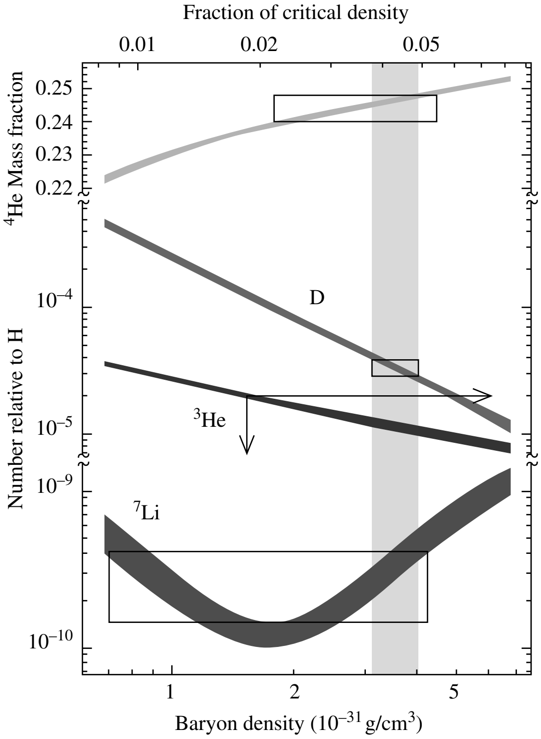

Fig. 8.3 The abundance of light nuclear elements vs. baryon mass density ρ_B of the universe. (Graph from Freedman and Turner (2003).) The curves are big bang nucleosynthesis predictions and the boxes are observational results: the vertical heights represent uncertainties in observation and horizontal width the range of ρ_B that theory can accommodate observation. The shaded vertical column represents the value of ρ_B that allows theory and observation to agree for all four elements. Its uncertainty (the width of the column) is basically determined by the deuterium abundance, which has both a strong ρ_B dependence and a well-measured value.

detailed information about the history and composition of the universe. This is a major tool for quantitative cosmology.

8.5.1 Universe became transparent to photons

The epoch when charged nuclear ions and electrons were transformed into neutral atoms is called the **photon-decoupling time**[11] (t_γ). This took place when the thermal energy of photons just dropped below the threshold required to ionize the newly formed atoms. Namely, the dominant reversible reaction during the age of ions,

$$e^- + H^+ \longleftrightarrow H + \gamma, \tag{8.60}$$

ceased to proceed from right to left when the photon energy was less than the ionization energy. All the charged electrons and ions were swept up and bound themselves into stable neutral atoms.

One would naturally expect the temperature at the decoupling time $k_B T_\gamma = O(\text{eV})$ to be comparable to the typical atomic binding energy. In fact, a detailed calculation yields

$$k_B T_\gamma \simeq 0.26 \text{ eV}. \tag{8.61}$$

Dividing out the Boltzmann's constant k_B, this energy corresponds to a photon temperature of

$$T_\gamma \simeq 3{,}000 \text{ K}. \tag{8.62}$$

The same $\Gamma = \text{H}$ calculation also yields the redshift

$$z_\gamma \simeq 1{,}100.$$

[11] The photon-decoupling time is also referred to in the literature as the "recombination time." We prefer not to use this terminology as, up to this time, ions and electrons had never been combined. The name has been used because of the analogous situation in the interstellar plasma where such atomic formation is indeed a re-combination.

If we know the matter and energy content of the universe now, we can translate this redshift value into an estimate of the cosmic age when photons decoupled. To do such a calculation properly we still need the material to be discussed in Chapter 9. This redshift can be translated into a cosmic age of

$$t_\gamma \simeq 10^{13}\ \text{s}, \tag{8.63}$$

that is, about 350,000 years after the big bang.

Shortly after recombination, the universe became transparent to the electromagnetic radiation. Thereafter, the decoupled photons could travel freely through the universe, but they still had the blackbody spectrum which was unchanged as the universe expanded. These relic photons cooled according to the scaling law of $T \propto a^{-1}$. Thus, the big bang cosmology predicted that everywhere in the present universe there should be a sea of primordial photons following a blackbody spectrum.

What should the photon temperature be now? From the estimates of $T_\gamma \simeq$ 3,000 K and $z_\gamma \simeq$ 1,100 we can use (8.40) and (7.54) to deduce

$$T_{\gamma,0} = \frac{T_\gamma}{1+z_\gamma} \simeq 2.7\ \text{K}.$$

A blackbody spectrum of temperature T has its maximal intensity at the wavelength $\lambda_{\max} T \simeq 0.3$ cm K (known as the "Wien displacement constant"). Thus, an estimate $T_{\gamma,0} \simeq 2.7$ K implies a thermal spectrum with the maximal intensity at $\lambda_{\max} = O(\text{mm})$—there should be a relic background radiation in the microwave range.[12]

In summary, photons in the early universe were tightly coupled to ionized matter through Thomson scattering. Such interactions stopped about 350,000 years after the big bang, when the universe had cooled sufficiently to form neutral atoms (mainly hydrogens). Ever since this last scattering time, the photons have traveled freely through the universe, and redshifted to microwave frequencies as the universe expanded. This primordial light should appear today as the CMB thermal radiation with a temperature of about 3 K.

8.5.2 The discovery of CMB radiation

The observational discovery of the CMB radiation was one of the great scientific events in the modern era. It made the big bang cosmology much more credible as it is difficult to see how else such a thermal radiation could have been produced. The discovery and its interpretation also constitute an interesting story. In 1964, Robert Dicke led a research group (including James Peebles, Peter Roll, and David Wilkinson) at Princeton University to detect this cosmic background radiation. While they were constructing their apparatus, Dicke was contacted by Arno Penzias and Robert W. Wilson at the nearby Bell Lab. Penzias and Wilson had used a horn-shaped microwave antenna in the past year to do astronomical observations of the Galaxy. This "Dicke radiometer" was originally used in a trial satellite communication experiment, and was known to have some "excess noise." Not content to ignore it, they made a careful measurement of this background radiation, finding it to be independent of direction, time of the day, or season of the year. While puzzling over the cause of such a radiation, they were informed by one of their colleagues of Princeton group's interest in the detection of a cosmic background radiation. (Peebles had given

[12] While this electromagnetic radiation is outside the visible range, we can still "see" it because such a microwave noise constitutes a percent of the television "snow" between channels.

a colloquium on this subject.) This resulted in the simultaneous publication of two papers: one in which Penzias and Wilson announced their discovery; the other by the Princeton group explaining the cosmological significance of the discovery. (At about the same time, a research group around Yakov Zel'dovich in Moscow also recognized the importance of the cosmic background radiation as the relic signature of a big bang beginning of the universe.)

Because of microwave absorption by water molecules in the atmosphere, it is desirable to carry out CMB measurements at locations having low humidity and/or at high altitude. Thus, some of the subsequent observations were done with balloon-borne instruments launched in Antarctica (low temperature, low humidity, and high altitude). Or even better, to go above the atmosphere in a satellite. This was first accomplished in the early 1990s (Smoot *et al.*, 1990) by the Cosmic Background Explorer satellite (COBE), obtaining results showing that the CMB radiation followed a perfect blackbody spectrum (Fixsen *et al.*, 1996) with a temperature of

$$T_{\gamma,0} = 2.725 \pm 0.002 \text{ K}. \tag{8.64}$$

The COBE observation not only confirmed that the thermal nature of the cosmic radiation was very uniform (the same temperature in every direction), but also discovered the minute anisotropy at the micro-Kelvin level. This has been interpreted as resulting from the matter density perturbation, which, through subsequent gravitational clumping, gave rise to the cosmic structure we see today: galaxies, clusters of galaxies, and voids, etc. This will be further discussed in Section 8.5.4, and in Chapter 9.

8.5.3 Photons, neutrinos, and the radiation–matter equality time

The knowledge of CMB's temperature allows us to calculate the photon number density. This reveals that there are about a billion photons to every nucleon in the universe. Such information will enable us to estimate the cosmic time when the universe made the transition from a radiation-dominated to a matter-dominated universe. In this section, we also discuss another cosmic thermal relic: the primordial neutrinos.

The photon to baryon number ratio

Knowing the CMB photon temperature $T_{\gamma,0} = 2.725$ K, we can calculate the relic photon number density via (8.35):

$$n_{\gamma,0} = \frac{2.4}{\pi^2}\left(\frac{k_\mathrm{B}T_{\gamma,0}}{\hbar c}\right)^3 \simeq 411/\text{cm}^3. \tag{8.65}$$

Namely, there are now in the universe, on an average, 400 photons for every cubic centimeter. Clearly this density is much higher than the baryon matter density obtained in (8.59) from the primordial nucleosynthesis theory, and the observed abundance of light nuclear elements. The baryon and photon number ratio comes out to be

$$\frac{n_\mathrm{B}}{n_\gamma} \simeq 0.6 \times 10^{-9}. \tag{8.66}$$

For every proton or neutron there are about 2 billion photons. This also explains why the thermal energy at photon decoupling is as low as 0.26 eV. Considering that the hydrogen ionization energy is 13.6 eV, why was the ionization not shut off until the thermal energy fell so much below this value? This just reflects the fact that there are so many photons for every baryon that the blackbody thermal photons have a broad distribution and photon number density was very high, $n \sim T^3$. Thus, even though the average photon energy was only 0.26 eV, there was a sufficient number of high energy photons at the tail end of the distribution to bring about a new equilibrium phase.

This ratio (8.66) should hold all the way back to the photon decoupling time, because not only was the number of free photons unchanged, but also the baryon number, since all the interactions in this low energy range respected the law of baryon number conservation. The relevance of this density ratio to the question of matter–antimatter asymmetry of the universe is discussed in Box 8.2.

Transition from radiation-dominated to matter-dominated era

It is clear that we now live in a matter-dominated universe, as the matter energy density is considerably greater than that for the radiation, $\Omega_{\rm M} \gg \Omega_{\rm R}$, where the radiation is composed of CMB photons[13] and three flavors of neutrinos.[14] Their relative abundance is calculated in Box 8.1: using (8.76) $\Omega_{\nu,0} = 0.68\ \Omega_{\gamma,0}$ so that the radiation density is about 1.68 times larger than the density due to CMB radiation alone. The matter–radiation ratio now can be related to the baryon–photon ratio

$$\begin{aligned}\frac{\Omega_{\rm M,0}}{\Omega_{\rm R,0}} &= \frac{\Omega_{\rm M,0}}{\Omega_{\rm B,0}} \frac{\Omega_{\rm B,0}}{1.68 \times \Omega_{\gamma,0}} \\ &= \frac{0.3}{0.04} \frac{n_{\rm B}}{n_\gamma} \frac{m_{\rm N}c^2}{1.68 \times k_{\rm B}T_{\gamma,0}} \simeq 1.1 \times 10^4, \end{aligned} \tag{8.67}$$

where the energy density for nonrelativistic baryon matter is given by the product of number density and rest energy of the nucleon $m_{\rm N}c^2 = 939$ MeV and photon energy by $k_{\rm B}T_{\gamma,0}$. We have also used the phenomenological results of (7.29) and (7.30) for $\Omega_{\rm M,0}$ and $\Omega_{\rm B,0}$. Since radiation density scales as $\rho_{\rm R} \sim a^{-4}$ while matter $\rho_{\rm M} \sim a^{-3}$, even though $\Omega_{\rm R,0}$ is small in the present epoch, radiation was the dominant contributor in the early universe. The epoch when the universe made this transition from a radiation-dominated to matter-dominated era can be fixed by the condition of $\rho_{\rm R}(t_{\rm RM}) = \rho_{\rm M}(t_{\rm RM})$, where $t_{\rm RM}$ is the cosmic age when radiation and matter densities were equal:

$$1 = \frac{\rho_{\rm R}}{\rho_{\rm M}} = \frac{\rho_{\rm R,0}}{\rho_{\rm M,0}}[a(t_{\rm RM})]^{-1} = \frac{\Omega_{\rm R,0}}{\Omega_{\rm M,0}}(1 + z_{\rm RM}) \simeq \frac{1 + z_{\rm RM}}{1.1 \times 10^4}. \tag{8.68}$$

Hence the redshift for radiation–matter equality is $z_{\rm RM} \simeq 1.1 \times 10^4$, which is 10 times larger than the photon decoupling time with $z_\gamma \simeq 1{,}100$. This also yields scale factor and temperature ratios of $[a(t_\gamma)/a(t_{\rm RM})] = [T_{\rm RM}/T_\gamma] \simeq 10$, or a radiation thermal energy

$$k_{\rm B}T_{\rm RM} = 10\ k_{\rm B}T_\gamma = O(10 \text{ eV}). \tag{8.69}$$

Knowing this temperature ratio we can find the "radiation–matter equality time" $t_{\rm RM} \simeq 10{,}000$ years (Problem 8.9) from the photon decoupling time $t_\gamma \simeq 10^{13}$ s.

[13] The energy density due to star light (i.e. all electromagnetic radiation except the microwave background) is much smaller than CMB photons.

[14] Even though we have not measured all the neutrino masses, all indications are that they are very small and we can treat them as relativistic particles with kinetic energy much larger than their rest energy.

From that time on, gravity (less opposed by significant radiation pressure) began to grow from the tiny lumpiness in matter distribution into the rich cosmic structures we see today.

Also, since radiation dominance ceased so long ago ($t_{\mathrm{RM}} \ll t_0$) and if the universe is composed of matter and radiation only, we can estimate the age of the universe based on the model of a matter-dominated universe because during the overwhelming part of the universe's history, the dominant energy component has been in the form of nonrelativistic matter. In particular, for a matter-dominated universe with a spatially flat geometry, we have, according to (8.30),

$$(t_0)^{k=0} \simeq (t_0)^{k=0}_{\mathrm{MDU}} = \frac{2}{3} t_{\mathrm{H}} \simeq 9 \text{ Gyr}. \tag{8.70}$$

This value is seen to be significantly less than the age deduced by observation (cf. Section 7.1.3). As we shall see in the next chapter, a flat geometry is indeed favored theoretically and confirmed by observation. This contradiction in the estimate of the cosmic age hinted at the possibility that, besides radiation and matter, there may be some other significant form of energy in the universe.

Box 8.1 Entropy conservation, photon reheating, and the neutrino temperature

We have already noted that (8.3), being a linear combination of the Friedmann Eqs (8.1) and (8.2), has the interpretation of energy conservation, $dE + pdV = 0$. This implies, through the Second Law of thermodynamics $TdS = dE + pdV$, that the entropy is conserved, $dS = 0$. Holding the volume fixed, a change in entropy is related that of energy density $dS = (V/T)du_{\mathrm{R}}$ and the radiation energy density u_{R} being proportional to T^4 as in (8.36), we can relate entropy to temperature and volume as

$$S = \frac{g^*}{2} a_{\mathrm{SB}} V \int \frac{dT^4}{T} = \frac{2g^*}{3} a_{\mathrm{SB}} V T^3. \tag{8.71}$$

Before neutrino decoupling, the radiation particles–photons, neutrinos, anti-neutrinos, electrons, and positrons–were in thermal equilibrium. Thus, photons and neutrinos (as well as anti-neutrinos) have the same temperature $T_\gamma = T_\nu$. We have already discussed the sequence of cosmic equilibria: first, neutrinos decoupled at $k_{\mathrm{B}} T \approx 3$ MeV, second there was photon decoupling at $k_{\mathrm{B}} T \approx O$ (eV). Whether they were coupled or not, the radiation temperature scaled as $T \propto a^{-1}$. So we would expect the relic photons and neutrinos to have the same temperature now. However, we must take into account the effect of positron disappearance, which happened at $k_{\mathrm{B}} T \approx O$ (MeV) in between these two epochs of neutrino and photon decoupling. The annihilation of electron and positron into photons would heat up the photons ("**photon reheating**"), raising the photon temperature over that of neutrinos, which had already decoupled. One can calculate the raised photon temperature $T'_\gamma > T_\gamma$ by the condition of entropy conservation discussed above:

$$S_\gamma + S_{\mathrm{e}^-} + S_{\mathrm{e}^+} = S'_\gamma. \tag{8.72}$$

With the effective spin degrees of freedom (8.38) in (8.71), this entropy conservation equation becomes $[2+\frac{7}{8}(2+2)]T^3V = 2T'^3V'$, or

$$\left(\frac{T'_\gamma}{T_\gamma}\right)^3 \frac{V'}{V} = \frac{11}{4}. \tag{8.73}$$

On the other hand, the neutrinos being noninteracting, their entropy was not affected, $S_\nu = S'_\nu$. Namely, $T_\nu^3 V = T_\nu'^3 V'$, or

$$\left(\frac{T'_\nu}{T_\nu}\right)^3 \frac{V'}{V} = 1. \tag{8.74}$$

Equations (8.73) and (8.74), together with the thermal equilibrium condition $T_\gamma = T_\nu$ that prevailed before the positron disappearance, lead to

$$T'_\gamma = \left(\frac{11}{4}\right)^{1/3} T'_\nu = 1.4 \times T'_\nu. \tag{8.75}$$

Knowing $T_{\gamma,0} \simeq 2.7$ K the temperature of relic neutrinos and antineutrinos now should be $T_{\nu,0} \simeq 1.9$ K. This temperature difference leads to the different photon and neutrino number densities as first stated in Section 8.3.2.

Using (8.75) we can also compare the neutrino and photon contributions to the radiation energy content of the universe via (8.36):

$$\frac{\rho_\nu}{\rho_\gamma} = \frac{g_\nu^*}{g_\gamma}\left(\frac{T'_\nu}{T'_\gamma}\right)^4 = \frac{21/4}{2}\left(\frac{4}{11}\right)^{4/3} = 0.68, \tag{8.76}$$

because the neutrino effective spin degrees of freedom $g_\nu^* = \frac{7}{8}[3\times(1+1)]$ must include the three species (“flavors”) of neutrinos (electron, muon, and tau neutrinos), as well as their antiparticles.

Box 8.2 Cosmological asymmetry between matter and antimatter

The ratio displayed in (8.66) is also a measure of the baryon number asymmetry in the universe. By this we mean that if the universe contained equal numbers of baryons (having baryon number +1) and antibaryons (having baryon number −1) then the net baryon number would vanish, $n_\mathrm{B} = 0$. In the early universe there was plenty of thermal energy so that anti-baryons (particles carrying negative baryon numbers such as anti-quarks making up anti-protons, etc.) were present in abundance. In fact, there were just about an equal number of particles and antiparticles. The fact that the present universe has only particles and no antiparticles means that there must have been a slight excess of particles. As the universe cooled to the point when particle–antiparticle pairs could no longer be produced by radiation, all antiparticles would disappear through annihilation. (See Section 8.3.2 on positron disappearance.) Thermal equilibrium in the early universe would ensure that photons and quark–anti-quark pair numbers should be comparable. From this we can conclude that the population of baryons was only slightly larger than that of anti-baryons as indicated, for example, by the quark–anti-quark asymmetry ratio:

$$\frac{n_q - n_{\bar{q}}}{n_q + n_{\bar{q}}} \backsimeq \frac{n_{\mathrm{B}}}{n_\gamma} = O(10^{-9}). \qquad (8.77)$$

Because the universe is observed to be electrically neutral, this statement about baryon number asymmetry can also be extended to electron–positron asymmetry. Namely, (8.77) holds for the entire matter–antimatter asymmetry of the universe.

This matter–antimatter asymmetry is a puzzle, as no cosmological model can generate a net baryon number if all underlying physical processes (such as all the interactions included in the Standard Model of particle physics) conserve baryon number. Thus one had to impose on the standard big bang cosmology an ad hoc asymmetric initial boundary condition (8.66). Why should there be this asymmetry, with this particular value? It would be much more satisfying if starting with a symmetric state (or better, independent of initial conditions) such an asymmetry could be generated by the underlying physical interactions. One of the attractive features of the Standard Model of particle interaction and its natural extensions is that they generally possess precisely the conditions to produce such excess of matter over antimatter.

8.5.4 CMB temperature fluctuation

After subtracting off the Milky Way foreground radiation, one obtained, in every direction, the same blackbody temperature—the CMB showed a high degree of isotropy. However, such an isotropy is not perfect. One of the major achievements of COBE satellite observations was the detection of slight variation of temperature: first at the 10^{-3} level associated with the motion of our Local Group of galaxies in the gravitational potential due to neighboring cosmic matter distribution, then at the 10^{-5} level, which, as we shall explain, holds the key to our understanding of the origin of structure in the universe, how the primordial plasma evolved into stars, galaxies, and clusters of galaxies.

The dipole anisotropy

The sensitive instrument Differential Microwave Radiometer (DMR) aboard COBE first revealed the existence of a "dipole anisotropy" in the CMB background. Although each point on the sky has a blackbody spectrum, in one half of the sky the spectrum corresponds to a slightly higher temperature while the other half is slightly lower with respect to the average background temperature

$$\frac{\delta T}{T} \approx 1.237 \times 10^{-3} \cos\theta, \qquad (8.78)$$

where θ is measured from the direction joining the hottest to the coldest spot on the sky. The dipole distortion is a simple Doppler shift, caused by the net motion of the COBE satellite which is 371 km/s relative to the reference frame in which the CMB is isotropic (cf. Problem 8.14). The Doppler effect changes the observed frequency, which in turn changes the energy and temperature of the detected background radiation. The different peculiar motions result from the gravitational attraction as a consequence of uneven distribution of masses in our cosmic neighborhood. The quoted number represents the observation result after subtracting out the orbit motion of COBE around the earth ($\sim$8 km/s), and the seasonal motion of earth around the sun ($\sim$30 km/s). The measured value is in fact the vector sum of the orbital motion of the solar system around the galactic center ($\sim$220 km/s), the motion of the Milky Way around the center of mass of the Local Group ($\sim$80 km/s), and the motion of Local Group of galaxies (630 ± 20 km/s) in the general direction of the constellation Hydra. The last, being the peculiar motion of our small galaxy cluster toward the large mass concentration in the neighboring part of the universe, reflects the gravitational attraction by the very massive Virgo cluster at the center of our Local Supercluster, which is in turn accelerating toward the Hydra–Centaurus supercluster.

The peculiar motions mentioned above are measured with respect to the frame in which the CMB is isotropic. The existence of such a CMB rest frame does not contradict special relativity. SR only says that no internal physical measurements can detect absolute motion. Namely, physics laws do not single out an absolute rest frame. It does not say that we cannot compare motion relative to a cosmic structure such as the microwave background. The more relevant question is why constant velocity motion in this CMB rest frame coincides with the Galilean frames of Newton's first law. (CMB acts as an aether.) To the extent that the CMB frame represents the average mass distribution of the universe, this statement is called **Mach's principle** (cf. Box 1.1). While to a large extent Einstein's GR embodies Mach's principle, there is no definitive explanation of why the CMB rest frame defines the inertial frames for us.

Physical origin and mathematical description of CMB anisotropy

After taking off this 10^{-3} level dipole anisotropy, the background radiation is seen to be isotropic. CMB being a snapshot of our universe, the observed isotropy is a direct evidence of our working hypothesis of a homogeneous and isotropic universe. Nevertheless, this isotropy should not be perfect. The observed universe has all sorts of structure, some of the superclusters of galaxies and largest voids have dimensions as large as 100 Mpc across. Such a basic feature of our universe must be reflected in the CMB in the form of small temperature anisotropy. There must be matter density nonuniformity which would have brought about temperature anisotropy through electromagnetic interactions; photons traveling from denser regions were gravitational redshifted and therefore arrived cooler, while photons from less dense regions did less work and arrived warmer. One of the great achievements of the COBE observation team is the first observation of such an anisotropy, at the level of 10^{-5} (Smoot *et al.*, 1992). Small temperature variations of about 10 μK coming from different directions was finally detected. This discovery provided the first evidence for a primordial density nonuniformity that, under gravitational attraction, grew into the structures of stars, galaxies, and clusters of galaxies that we observe today. $\delta T/T = O(10^{-5})$ was smaller than expected based on the observed structure of luminous matter. But this "discrepancy" can be resolved by the existence of exotic dark matter. There were further inhomogeneities not seen through the CMB anisotropy because they were due to matter that did not have electromagnetic interactions to leave an imprint on the background photons.

Here we present the basic formalism needed for a description of this CMB anisotropy. The CMB temperature has directional dependence $T(\theta, \phi)$ with an average of

$$\langle T \rangle = \frac{1}{4\pi} \int T(\theta, \phi) \sin\theta \, d\theta \, d\phi = 2.725 \text{ K}. \tag{8.79}$$

The temperature fluctuation

$$\frac{\delta T}{T}(\theta, \phi) \equiv \frac{T(\theta, \phi) - \langle T \rangle}{\langle T \rangle} \tag{8.80}$$

has a root-mean-square value of

$$\left\langle \left(\frac{\delta T}{T} \right)^2 \right\rangle^{1/2} = 1.1 \times 10^{-5}. \tag{8.81}$$

How do we describe such temperature variation across the celestial sphere? Recall that for a function of one variable, a useful approach is Fourier expansion of the function in a series of sine waves with frequencies that are integral multiples of the fundamental wave (with the largest wavelength). Similarly for the dependence on (θ, ϕ) by the temperature fluctuation (think of it as vibration modes on the surface of an elastic sphere), we expand it in terms of spherical harmonics[15]

$$\frac{\delta T}{T}(\theta, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} a_{lm} Y_l^m(\theta, \phi). \tag{8.82}$$

These basis functions obey the orthonormality condition

$$\int Y_l^{*m} Y_{l'}^{m'} \sin\theta d\theta d\phi = \delta_{ll'} \delta_{mm'} \tag{8.83}$$

and the addition theorem

$$\sum_m Y_l^{*m}(\hat{n}_1) Y_l^m(\hat{n}_2) = \frac{2l+1}{4\pi} P_l(\cos\theta_{12}), \tag{8.84}$$

where $P_l(\cos\theta)$ is the Legendre polynomial, $\hat{n}_1$ and $\hat{n}_2$ are two unit vectors pointing in directions with an angular separation θ_{12}. Namely, $\hat{n}_1 \cdot \hat{n}_2 = \cos\theta_{12}$. We display a few samples of the spherical harmonics,

$$Y_0^0 = \left(\frac{1}{4\pi} \right)^{1/2}, \quad Y_1^0 = \left(\frac{3}{4\pi} \right)^{1/2} \cos\theta,$$

$$Y_3^{\pm 1} = \mp \left(\frac{21}{64\pi} \right)^{1/2} \sin\theta (5\cos^2\theta - 1) e^{\pm i\phi}.$$

The multipole number "l" represents the number of nodes (locations of zero amplitude) between equator and poles, while "m" is the longitudinal node number. For a given l, there are $2l + 1$ values for m: $-l, -l+1, \ldots, l-1, l$. The expansion coefficients a_{lm} are much like the individual amplitudes in a Fourier series. They can be projected out from the temperature fluctuation by (8.83):

$$a_{lm} = \int Y_l^{*m}(\theta, \phi) \frac{\delta T}{T}(\theta, \phi) \sin\theta d\theta d\phi. \tag{8.85}$$

Cosmological theories predict only statistical information. The most useful statistics is the 2-point correlation. Consider two points at $\hat{n}_1$ and $\hat{n}_2$ separated

[15] The temperature being real, the expansion could equally be written in terms of $a_{lm}^* Y_l^{*m}$.

by θ. We define the correlation function

$$C(\theta) \equiv \left\langle \frac{\delta T}{T}(\hat{n}_1)\frac{\delta T}{T}(\hat{n}_2) \right\rangle_{\hat{n}_1\cdot\hat{n}_2=\cos\theta}, \tag{8.86}$$

where the angle brackets denote the averaging over an ensemble of realizations of the fluctuation.[16] The inflationary cosmology predicts that the fluctuation is Gaussian as is thus independent of the a_{lm}s. Namely, the multipoles a_{lm} are uncorrelated for different values of l and m:

$$\langle a^*_{lm}a_{l'm'} \rangle = C_l\delta_{ll'}\delta_{mm'}, \tag{8.87}$$

which defines the power spectrum C_l as a measure of the relative strength of spherical harmonics in the decomposition of the temperature fluctuations. The lack of m-dependence reflects the rotational symmetry of the underlying cosmological model. When we plug (8.82) into (8.86), the conditions (8.87) and (8.84) simplify the expansion to

$$C(\theta) = \frac{1}{4\pi}\sum_{l=0}^{\infty}(2l+1)C_lP_l(\cos\theta). \tag{8.88}$$

Namely, the information carried by $C(\theta)$ in the angular space can be represented by C_l in the space of multipole number l. The power spectrum C_l is the focus of experimental comparison with theoretical predictions. From the map of measured temperature fluctuation, one can extract multipole moments by the projection (8.85) and since we do not actually have an ensemble of universes to take the statistical average, this is estimated by averaging over a_{lm}s with different ms. Such an estimate will be uncertain by an amount inversely proportional to the square-root of the number of samples

$$\left\langle \left(\frac{\delta C_l}{C_l}\right)^2 \right\rangle^{1/2} \propto \sqrt{\frac{1}{2l+1}}. \tag{8.89}$$

The expression also makes it clear that the variance will be quite significant for low multiple moments when we have only a very small number of samples. This is referred to as the "cosmic variance problem" (cf. Fig. 9.13).

In the next chapter we shall present the basic features of the power spectrum: to show how it can be used to measure the curvature of space, to test different theories of the origin of the cosmic structure that we see today, and to extract many cosmological parameters.

[16] In principle it means averaging over many universes. Since we have only one universe, this ensemble averaging is carried out by averaging over multiple moments with different m moments, which in theory should be equal because of spherical symmetry.

Review questions

1. Describe the relation of the Friedmann Eq. (8.1) and the Einstein equation, as well as give its Newtonian interpretation. Why can we use non-relativistic Newtonian theory to interpret the general relativistic equation in cosmology? Also, in what sense is it only quasi-Newtonian?
2. In what sense can the critical density be understood as akin to the more familiar concept of escape velocity?
3. Why do we expect the energy density of radiation to scale as a^{-4}? Why should the energy of the universe be radiation-dominated in its earliest moments?

4. What is the equation of state parameter w for radiation? What is the time dependence of the scale factor $a(t)$ in a flat radiation-dominated universe (RDU) and in a flat matter-dominated universe (MDU)? How is the age of the universe t_0 related to the Hubble time $t_{\rm H}$ in a RDU, and in a MDU? Justify the natural expectation that the age of our universe is approximately two-thirds of the Hubble time.

5. Draw a schematic diagram showing the behaviors of the scale factor $a(t)$ for various values of k in cosmological models (with zero cosmological constant). (It is suggested that all $a(t)$ curves be drawn to meet at the same point $a(t_0)$ with the same slope $\dot{a}(t_0)$). Also mark the regions corresponding to a decelerating universe, an accelerating universe, and an empty universe.

6. Give an argument for the scaling behavior of the radiation temperature: $T \backsim a^{-1}$. Show that under such a scaling law, the spectrum distribution of the blackbody radiation is unchanged as the universe expands.

7. What is the condition (called the Gamow condition) for any particular set of interacting particles being in thermal equilibrium during the various epochs of the expanding universe?

8. Given that the cosmic helium synthesis took place when the average thermal energy of particles was of the order of MeV, how would you go about estimating the number density ratio of neutron to proton $n_{\rm n}/n_{\rm p}$ at that epoch? If $n_{\rm n}/n_{\rm p} \backsimeq \frac{1}{7}$, what is the cosmic helium **mass fraction**?

9. How can one use the theory of big bang nucleosynthesis and the observed abundance of light elements such as deuterium to deduce the baryon number density $\Omega_{\rm B}$ and that the number of neutrino flavors should be three?

10. What physics process took place around the photon decoupling time t_γ? What were the average thermal energy and temperature at t_γ? Knowing the redshift $z_\gamma \simeq 10^3$, calculate the expected photon temperature now.

11. What is the cosmic time when the universe made the transition from a radiation-dominated to a matter-dominated system. How does it compare to the nucleosynthesis time, and photon decoupling times?

12. Give the argument that relates the matter–antimatter asymmetry in the early universe to the baryon-to-photon ratio now ($\simeq 10^{-9}$).

13. Why would the peculiar motion of our galaxy show up as CMB dipole anisotropy?

14. Besides the dipole anisotropy, how does the CMB temperature anisotropy reflect the origin of cosmic structure?

15. What is the “cosmic variance”?

Problems

(8.1) **Friedmann equations and energy conservation** Show that a linear combination of these two Friedmann Eqs (8.1) and (8.2) leads to Eq. (8.3).

(8.2) **Newtonian interpretation of the second Friedmann equation** Adopting the same approach used in the Newtonian “derivation” of Eq. (8.1) in the text, interpret the second Friedmann Eq. (8.2) as the $F = ma$ equation of the system.

(8.3) **Friedmann equation for a multi-component universe** Show that the Friedmann equation for a multi-component universe may be written as

$$\dot{a}^2 + \frac{kc^2}{R_0^2} = \frac{8\pi G_{\rm N}}{3} \sum_w \rho_{w,0} a^{-(1+3w)},$$

where w is the equation of state parameter defined in (8.4).

(8.4) **The empty universe** A low density universe may be approximated by setting the density function in the Friedmann equation to zero,

$$\dot{a}^2 = -\frac{kc^2}{R_0^2}.$$

Besides the uninteresting possibility of $\dot{a} = k = 0$ for a static universe with a Minkowski spacetime, show that the nontrivial solution to this equation is represented by the straight-line $a(t)$ in Fig. 8.2. Find the Hubble relation between the proper distance and redshift in such a model universe.

(8.5) **Hubble plot in a matter-dominated flat universe** As we have explained at the end of Chapter 7, the Hubble diagram is usually a plot of the distance modulus vs. redshift. Find this relation for a matter-dominated universe with a flat spatial geometry $k = 0$.

(8.6) **Another calculation of photon density** Give a direct estimate of thermal photon number density from the estimate that at a cosmic era with redshift $z_\gamma \simeq 1{,}100$ the average photon energy was $\bar{u} \simeq 0.26$ eV.

(8.7) **Time and redshift of a light emitter** Given the time dependence of the scale factor as in (8.27) $a(t) = (t/t_0)^x$, use (7.49) to calculate the proper distance $d_\mathrm{p}(t)$ between a light emitter (at t_em) and receiver (at t_0) in terms of the emission time t_em as well as another expression in terms of its redshift z.

(8.8) **Scaling behavior of number density and Hubble's constant**

(a) Show that the number densities for matter and radiation both scale as

$$\frac{n(t)}{n_0} = (1+z)^3$$

with the redshift.

(b) From the Friedmann equation, show that the Hubble constant $H(t)$ scales as

$$H^2 = \Omega_{\mathrm{M},0}(1+z)^3 H_0^2$$

in a matter-dominated flat universe, and as

$$H^2 = \Omega_{\mathrm{R},0}(1+z)^4 H_0^2$$

in a radiation-dominated flat universe.

(8.9) **Radiation and matter equality time** Knowing that the photon decoupling epoch corresponds to a redshift of $z_\gamma = 1.1 \times 10^3$ and a cosmic time $t_\gamma \simeq 350{,}000$ year, convert the radiation–matter equality redshift $z_\mathrm{RM} \simeq 1.1 \times 10^4$ from (8.68) to the corresponding cosmic time t_RM.

(8.10) **Density and deceleration parameter** In Problem 7.11 we introduce the deceleration parameter q_0. Use the second Friedmann equation (8.2) and the equation of state parameter w of (8.4) to show that

$$q_0 = \tfrac{1}{2}\sum_i \Omega_{i,0}(1+3w_i).$$

In particular in a matter-dominated flat universe $q_0 = +\frac{1}{2}$.

(8.11) **Temperature and redshift** Knowing how the temperature scales, show that we can also connect $T(z)$ at an epoch to the corresponding redshift z to T_0 at the present era:

$$T = T_0(1+z). \tag{8.90}$$

(8.12) **Radius of the universe** Show that the radius R_0 of the universe [cf. Eq. (7.40)] with $\Omega_0 > 1$, is related to the density parameter Ω_0 and the Hubble constant H_0 by

$$R_0 = \frac{c}{H_0\sqrt{\Omega_0 - 1}}.$$

(8.13) **Cosmological limit of neutrino mass** Given that the density parameter of non-baryonic dark matter $\Omega_\mathrm{exotic} = \Omega_\mathrm{M} - \Omega_\mathrm{B} \simeq 0.26$, what limit can be obtained for the average mass of neutrinos (average over three flavors)?

(8.14) **Temperature dipole anisotropy as Doppler effect** Show that the Doppler effect implies that an observer moving with a nonrelativistic velocity $\mathbf{v}$ through an isotropic CMB would see a temperature dipole anisotropy of

$$\frac{\delta T}{T}(\theta) = \frac{v}{c}\cos\theta,$$

where θ is angle from the direction of the motion.

Inflation and the accelerating universe

9

- Einstein introduced the cosmological constant in his field equation so as to obtain a static cosmic solution.
- The cosmological constant is the vacuum-energy of the universe: this constant energy density corresponds to a negative pressure, giving rise to a repulsive force that increases with distance. A vacuum-energy dominated universe expands exponentially.
- The inflationary theory of cosmic origin—that the universe had experienced a huge expansion at the earliest moment of the big bang—can provide the correct initial conditions for the standard model of cosmology: solving the flatness, horizon problems, and providing an origin of matter/energy, as well as giving just the right kind of density perturbation for subsequent structure formation.
- The inflationary epoch leaves behind a flat universe, which can be compatible with the observed matter density being less than the critical density and a cosmic age greater than 9 Gyr if there remains a small but nonvanishing cosmological constant—a dark energy. This would imply a universe now undergoing an accelerating expansion.
- The measurement of supernovae at high redshift provided direct evidence for an accelerating universe. Such data, together with other observational results, especially the anisotropy of cosmic microwave background (CMB) and large structure surveys, gave rise to a concordant cosmological picture of a spatially flat universe $\Omega = \Omega_\Lambda + \Omega_\mathrm{M} = 1$, dominated by dark energy $\Omega_\Lambda \approx 0.7$ and $\Omega_\mathrm{M} = \Omega_\mathrm{DM} + \Omega_\mathrm{LM} \approx 0.3$, and by dark matter $\Omega_\mathrm{DM} \gg \Omega_\mathrm{LM}$. The cosmic age comes out to be comparable to the Hubble time $t_0 \cong 13.2\,\mathrm{Gyr}$.
- The cosmological constant and the cosmic coincidence problems point to the need of new fundamental physics.

As we have discussed in Sections 7.1.3 and 8.2, Newton's and the original Einstein's equations would lead us to expect the expansion of the universe to slow down because of gravitational attraction. In this chapter, we shall see how a modification of the Einstein equation, with the introduction of the **cosmological constant** Λ, allows for the possibility of a gravitational repulsive force that increases with distance. This effect was first discovered by Einstein in his effort of seeking a static solution to the GR field equation. It also allows for the possibility that the universe had undergone an extraordinarily rapid expansion at an early moment (the inflationary epoch). The inflationary scenario of the big bang brings about just the correct initial conditions for the standard cosmology and predicts a flat geometry for the universe at large. Finally, a nonvanishing Λ term can account for the recently discovered evidence of an accelerating universe in the present epoch. An accelerating expansion means slower expansion in the past, hence a longer age for the universe—long enough to account for the oldest objects observed in the universe. The cosmological constant also provides us with a dark energy that, together with the observed matter content, fulfills the inflationary cosmology's prediction of a flat universe, which requires the mass/energy density of the universe to be equal to the critical density.

9.1 The cosmological constant

Before Hubble's discovery in 1929 of an expanding universe, just about everyone, Einstein included, believed that we lived in a static universe. Recall that the then-observed universe consisted essentially of stars within the Milky Way galaxy. But gravity, whether nonrelativistic or relativistic, is a universal attraction. Hence, theoretically speaking, a static universe is an impossibility. Specifically, as we have demonstrated, the Friedmann cosmological Eqs (8.1) and (8.2) have solutions corresponding always to a **dynamic** universe—a universe which is either contracting or expanding. Namely, these equations are not compatible with the static condition of an unchanging scale factor $\dot{a} = \ddot{a} = 0$, which would lead to a trivial empty universe,[1] $\rho = p = 0$.

Recall our brief discussion of the GR field equation $G_{\mu\nu} = \kappa T_{\mu\nu}$ with $\kappa = -8\pi c^{-4} G_{\mathrm{N}}$ in Section 5.3.2. $G_{\mu\nu}$ on the left-hand side (LHS) is the curvature tensor of spacetime and $T_{\mu\nu}$ on the right-hand side (RHS), the energy–momentum source term for gravity (the curved spacetime). The goal of obtaining a static universe from general relativity (GR) led Einstein to alter his field equation to make it contain a repulsion component. This could, in principle, balance the usual gravitational attraction to yield a static cosmic solution. Einstein discovered that the geometry side of his field equation can naturally accommodate an additional term. As will be discussed in Section 12.4.3, the simplest term that is mathematically compatible with Einstein's field Eq. (5.37) is the metric tensor $g_{\mu\nu}$,

$$G_{\mu\nu} - \Lambda g_{\mu\nu} = \kappa T_{\mu\nu}. \tag{9.1}$$

Such a modification will, however, alter its nonrelativistic limit to differ from Newton's equation. In order that this alteration is compatible with known phenomenology, it must have a coefficient Λ so small as to be unimportant in all situations except on truly large cosmic scales. Hence, this additional constant Λ has come to be called the **cosmological constant**.

While we have introduced this term as an additional geometric term, we could just as well move it to the RHS of the equation and view it as an additional source term of gravity. In particular, when the regular energy–momentum is absent $T_{\mu\nu} = 0$ (the vacuum state),

$$G_{\mu\nu} = \Lambda g_{\mu\nu} \equiv \kappa T^{\Lambda}_{\mu\nu}.$$

$T^{\Lambda}_{\mu\nu} = \kappa^{-1}\Lambda g_{\mu\nu} = (-c^4\Lambda/8\pi G_{\mathrm{N}})g_{\mu\nu}$ can be interpreted as the energy–momentum tensor of the vacuum. Just as $T_{\mu\nu}$ for cosmic fluid's ordinary radiation and matter depends on two functions of energy density ρ and pressure

[1] For Einstein equation without cosmological constant, a static solution necessarily corresponds to an empty universe. On the other hand, an empty universe is compatible with an expanding universe with negative spatial curvature. See Problem 8.4.

p, this vacuum-energy–momentum tensor $T^{\Lambda}_{\mu\nu}$ can be similarly parametrized by "**vacuum-energy density**" ρ_Λ and "**vacuum pressure**" p_Λ. As we shall demonstrate in Section 12.4.3 (after we have properly studied energy–momentum tensor in Section 10.3), these two quantities are related to a positive cosmological constant Λ as follows: the vacuum-energy per unit volume,

$$\rho_\Lambda = \frac{\Lambda c^2}{8\pi G_\mathrm{N}} > 0, \tag{9.2}$$

is a constant (in space and in time) and the corresponding vacuum pressure,

$$p_\Lambda = -\rho_\Lambda c^2 < 0, \tag{9.3}$$

is negative, corresponding to an equation-of-state parameter $w = -1$ as defined in Eq. (8.4). Such density and pressure, as we shall presently show, are compatible with basic physics principles, and, most relevant for our cosmological discussion, they give rise to a gravitational repulsion.

Λ as constant energy density and negative pressure

What is a negative pressure? Consider the simple case of a piston chamber filled with ordinary matter and energy, which exerts a positive pressure by pushing out against the piston. If it is filled with this Λ energy, Fig. 9.1, it will exert a negative pressure by **pulling in** the piston. Physically this is sensible because, as its energy per unit volume $\rho_\Lambda c^2$ is a constant, the change in system's energy is strictly proportional to its volume change $dE = \rho_\Lambda c^2 dV$. The system would like to lower its energy by volume-contraction (pulling in the piston). When we increase the volume of the chamber $dV > 0$ (hence its energy $dE > 0$) by pulling out the piston, we have to do positive work to overcome the pulling by the Λ energy. Energy conservation is maintained in such a situation because the negative pressure $p < 0$ is just what is required by the First Law of thermodynamics: $dE = -pdV$ when both dE and dV have the same sign. In fact the First Law also makes it clear that if energy density is a constant $dE = \rho c^2 dV$ so that the dV factors cancel, the pressure must necessarily equal the negative of the energy density $p = -\rho c^2$.

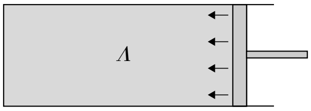

Fig. 9.1 The Λ energy in a chamber has negative pressure and thus pulls in the piston.

9.1.1 Vacuum-energy as source of gravitational repulsion

To see that the negative pressure can give rise to a repulsive force, let us first discuss the Newtonian limit of the Einstein equation with a general source, composed of mass density ρ as well as pressure p (as is the case for a cosmology with ideal fluid as the source). It can be shown (see Box 12.1 for detail), that the limiting equation, written in terms of the gravitational potential Φ, is

$$\nabla^2\Phi = 4\pi G_\mathrm{N}\left(\rho + 3\frac{p}{c^2}\right). \tag{9.4}$$

This informs us that not only mass, but also pressure, can be a source of gravitational field. For the nonrelativistic matter having a negligible pressure term, we recover the familiar Eq. (5.36) of Newton.

Explicitly displaying contributions from ordinary matter and vacuum-energy (thus density and pressure each have two parts: $\rho = \rho_\mathrm{M} + \rho_\Lambda$ and $p = p_\mathrm{M} + p_\Lambda$), the Newton/Poisson Eq. (9.4) becomes

$$\begin{aligned}\nabla^2\Phi &= 4\pi G_{\rm N}\left(\rho_{\rm M} + 3\frac{p_{\rm M}}{c^2} + \rho_\Lambda + 3\frac{p_\Lambda}{c^2}\right)\\ &= 4\pi G_{\rm N}\rho_{\rm M} - 8\pi G_{\rm N}\rho_\Lambda = 4\pi G_{\rm N}\rho_{\rm M} - \Lambda c^2,\end{aligned} \tag{9.5}$$

where we have used (9.3), $p_\Lambda = -\rho_\Lambda c^2$, and set $p_{\rm M} = 0$ because $\rho_{\rm M}c^2 \gg p_{\rm M}$. For the vacuum-energy dominated case of $\Lambda c^2 \gg 4\pi G_{\rm N}\rho_{\rm M}$, the Poisson equation can be solved (after setting the potential to zero at the origin) by

$$\Phi_\Lambda(r) = -\frac{\Lambda c^2}{6}r^2. \tag{9.6}$$

Between any two mass points, this potential corresponds to a repulsive force (per unit mass) that increases with separation **r**,

$$\mathbf{g}_\Lambda = -\nabla\Phi_\Lambda = +\frac{\Lambda c^2}{3}\mathbf{r}, \tag{9.7}$$

in contrast to the familiar $-\mathbf{r}/r^3$ gravitational attraction. With this pervasive repulsion that increases with distance, even a small Λ can have a significant effect on truly large dimensions. It would be possible to counteract the gravitational attraction and allow for the static solution sought by Einstein.

9.1.2 The static universe

We now consider the Friedmann Eqs (8.1) and (8.2) with a nonvanishing cosmological constant,

$$\frac{\dot a^2 + kc^2/R_0^2}{a^2} = \frac{8\pi G_{\rm N}}{3}(\rho_{\rm M} + \rho_\Lambda), \tag{9.8}$$

$$\frac{\ddot a}{a} = -\frac{4\pi G_{\rm N}}{c^2}\left[(p_{\rm M} + p_\Lambda) + \frac{1}{3}(\rho_{\rm M} + \rho_\Lambda)c^2\right]. \tag{9.9}$$

The RHS of (9.9) need not necessarily be negative because of the presence of the negative pressure term $p_\Lambda = -\rho_\Lambda c^2$. Consequently, a decelerating universe is no longer the inevitable outcome. For nonrelativistic matter, after setting $p_{\rm M} = 0$, we have

$$\frac{\ddot a}{a} = -\frac{4\pi G_{\rm N}}{3}\left(\rho_{\rm M} - 2\rho_\Lambda\right). \tag{9.10}$$

The static condition of $\ddot a = 0$ now leads to the constraint:

$$\rho_{\rm M} = 2\rho_\Lambda = \frac{\Lambda c^2}{4\pi G_{\rm N}}. \tag{9.11}$$

Namely, the mass density $\rho_{\rm M}$ of the universe is fixed by the cosmological constant. The other static condition of $\dot a = 0$ implies, through (9.8), the static solution $a = a_0 = 1$

$$\frac{kc^2}{R_0^2} = 8\pi G_\mathrm{N}\rho_\Lambda = \Lambda c^2. \tag{9.12}$$

Since the RHS is positive, we must have

$$k = +1. \tag{9.13}$$

Namely, the static universe must have a positive curvature (a closed universe) with a scale factor, the "radius of the universe," also being determined by the cosmological constant:

$$R_0 = \frac{1}{\sqrt{\Lambda}}. \tag{9.14}$$

Thus, the basic features of such a static universe, the density and radius, are determined by the arbitrary input parameter Λ. Not only is this a rather artificial arrangement, but also the solution is, in fact, unstable. Namely, a small variation will cause the universe to deviate from this static point. A slight increase in the separation will cause the gravitational attraction to decrease. Since the vacuum repulsion is not affected, the negative pressure prevails over the positive attraction, causing the separation to increase further. A slight decrease in the separation will increase the gravitational attraction to cause the separation to decrease further, until the whole system collapses.

Box 9.1 Some historical tidbits of modern cosmology

- The Friedmann equations with both ordinary and vacuum-energies (9.8) and (9.9) are sometimes call **the Friedmann–Lemaître equations**. That Einstein's equation had expanding, or contracting, solutions was first pointed out in the early 1920s by the Russian meteorologist and mathematician Alexander A. Friedmann, who also discovered Einstein's original oversight of not realizing his static solution being unstable. Friedmann's fundamental contribution to cosmology was hardly noticed by his contemporaries. It had to be rediscovered later by the Belgian civil engineer and priest Georges Lemaître, who published in 1927 his model of cosmology with contributions coming from both ρ_M and ρ_Λ. More importantly, Lemaître was the first one, having been aware of Hubble's work through his contact with Harvard astronomers (he spent three years studying at Cambridge University and M.I.T.), to show that the linear relation between distance and redshift (Hubble's law) follows from such cosmological considerations. The original derivations by Friedmann and Lemaître were somewhat awkward. Modern presentations have mainly followed the approach initiated by Howard Percy Robertson and Arthur G. Walker. Thus the framework using Einstein's equation for a homogeneous and isotropic universe has come to be known as the FLRW (Friedmann–Lemaître–Robertson–Walker) **cosmological model**.
- Having missed the chance of predicting an expanding universe before its discovery, Einstein came up with a solution which did not really solve the perceived difficulty. (His static solution is also unstable.)

It had often been said that later in life Einstein considered the introduction of the cosmological constant to be "the biggest blunder of his life!" This originated from a characterization by George Gamow in his autobiography (Gamow, 1970):

> Thus, Einstein's original gravity equation was correct, and changing it was a mistake. Much later, when I was discussing cosmological problems with Einstein, he remarked that the introduction of the cosmological term was the biggest blunder he ever made in his life.

Then Gamow went on to say,

> But this blunder, rejected by Einstein, is still sometimes used by cosmologists even today, and the cosmological constant Λ rears its ugly head again and again and again.

What we can conclude for sure is that Gamow himself considered the cosmological constant "ugly" (because this extra term made the field equation less simple). Generations of cosmologists continued to include it because there was no physical principle one could invoke to exclude this term. (If it is not forbidden, it must exist!) In fact, the discovery of the cosmological constant as the source of a new cosmic repulsive force must be regarded as one of Einstein's great achievements. Now, as we shall see, the idea of a nonzero cosmological constant was the key in solving a number of fundamental problems in cosmology. Namely, Einstein taught us the way to bring about a gravitational repulsion. Although this "tool" of the cosmological constant was not required for the task (the static universe) it was invented for, this repulsive force was needed to account for the explosion that was the big bang (inflationary epoch), and was needed to explain how the expansion of the universe could accelerate.

9.2 The inflationary epoch

The standard model of cosmology (the FLRW model) has been very successful in presenting a self-contained picture of the evolution and composition of the universe: how the universe expanded and cooled after the big bang; how the light nuclear elements were formed; after the inclusion of the proper density inhomogeneity, how in an expanding universe matter congealed to form stars, galaxies, and clusters of galaxies. It describes very well the aftermath of the big bang. However, the model says very little about the nature of the big bang itself: how did this "explosion of the space" come about? It assumes that all matter existed from the very beginning. Furthermore, it assumes certain very precise initial conditions that just cry out for an explanation (see the flatness and horizon problems discussed later).

The inflationary cosmology is an attempt to give an account of this big bang back to an extremely short instant (something like 10^{-38} s) after the $t = 0$ cosmic singularity.[2] According to this inflationary model, the universe had a burst of expansion during which the scale factor increased by more than

[2] This is to be compared to the even earlier period, comparable to the Planck time $t_{\mathrm{Pl}} = O(10^{-43}\ \mathrm{s})$, when quantum gravity is required for a proper description. See Section A.2.

30 orders of magnitude, see Fig. 9.2. In this inflationary process, all the matter and energy could have been created virtually from nothing. Afterwards, the universe followed the course of adiabatic expansion and cooling as described by the FLRW cosmology, presented in Chapter 8. Figure 9.2 also makes it clear that in the inflationary scenario, the observable universe originates from an entity some 10^{-30} times smaller than that which would have been the size in the case without inflation.

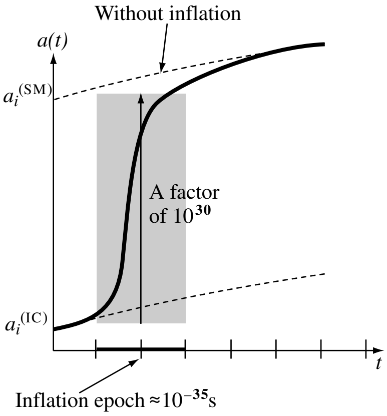

Fig. 9.2 Comparison of scale factor's time evolution. The standard FLRW model curves are represented by dashed lines; the solid curve is that of the inflation model which coincides with the standard model curve after 10^{-35} s. The intercepts on the a-axis correspond to the initial scales: $a_i^{(\text{SM})}$ in the standard model (without inflation) and $a_i^{(\text{IC})}$ in the inflation cosmology, respectively.

9.2.1 Initial conditions for the standard big bang model

The standard FLRW model requires a number of seemingly unnatural fine-tuned initial conditions. As we shall see, they are just the conditions that would follow from an inflationary epoch. We start the discussion of initial conditions by listing two such theoretical difficulties, two "problems."

The flatness problem

Because of gravitational attraction among matter and energy, we would expect the expansion of the universe to slow down. This deceleration $\ddot{a}(t) < 0$ means that $\dot{a}(t)$ must be a decreasing function. This is exemplified by the specific case of a radiation dominated universe $a \sim t^{1/2}$, thus $\dot{a} \sim t^{-1/2}$, or a matter dominated universe $a \sim t^{2/3}$, and $\dot{a} \sim t^{-1/3}$, as derived in (8.30). Recall that the Friedmann equation can be written in terms of the mass density parameter Ω as in (8.7):

$$[1 - \Omega(t)] = \frac{-kc^2}{[\dot{a}(t)]^2 R_0^2}. \tag{9.15}$$

This displays the connection between geometry and matter/energy: if $k = 0$ (a flat geometry), we must have the density ratio $\Omega = 1$ exactly; when $k \neq 0$ for an universe having curvature, then $[1 - \Omega(t)]$ must be **ever-increasing** because the denominator on the RHS is ever decreasing. Thus, the condition for a flat universe $\Omega = 1$ is an **unstable equilibrium point**—if Ω ever deviates from 1, this deviation will increase with time. Or, we may say: "gravitational attraction loves curvature"—it always enhances any initial curvature. In light of this property, it is puzzling that the present mass density Ω_0 has been found observationally (see Section 7.1.4) to be not too different from the critical density value $(1 - \Omega_0) = O(1)$. This means that Ω must have been extremely close to unity (extremely flat) in the cosmic past. Such a fine-tuned initial condition would require an explanation.

We can make such statement quantitatively. Ever since the radiation–matter equality time $t > t_{\text{RM}}$, with $z_{\text{RM}} = O(10^4)$ (cf. (8.68)) the evolution of the universe has been dominated by nonrelativistic matter: $a(t) \sim t^{2/3}$ or $\dot{a} \sim t^{-1/3} \sim a^{-1/2}$. We can then estimate the ratio as given by (9.15):

$$\frac{1 - \Omega(t_{\text{RM}})}{1 - \Omega(t_0)} = \left[\frac{\dot{a}(t_{\text{RM}})}{\dot{a}(t_0)}\right]^{-2} = \left[\frac{a_{\text{RM}}}{a_0}\right]$$
$$= (1 + z_{\text{RM}})^{-1} = O(10^{-4}). \tag{9.16}$$

Successful prediction of light element abundance by primordial nuclear synthesis gave us direct evidence for the validity of the standard model of cosmology back to the big bang nucleosynthesis time $t_{\text{bbn}} = O(10^2\,\text{s})$. The time evolution for $t < t_{\text{RM}}$ was radiation dominated: $a(t) \sim t^{1/2}$ or $\dot{a} \sim t^{-1/2} \sim a^{-1}$. This would then imply

$$\begin{aligned}\frac{1-\Omega(t_{\text{bbn}})}{1-\Omega(t_{\text{RM}})} &= \left[\frac{\dot{a}(t_{\text{bbn}})}{\dot{a}(t_{\text{RM}})}\right]^{-2} = \left[\frac{a(t_{\text{bbn}})}{a(t_{\text{RM}})}\right]^{2} \\ &= \left[\frac{k_{\text{B}}T_{\text{bbn}}}{k_{\text{B}}T_{\text{RM}}}\right]^{-2} \simeq O(10^{-10}), \qquad (9.17)\end{aligned}$$

where we have used the scaling behavior of the temperature, and (8.53) $k_{\text{B}}T_{\text{bbn}} = O(\text{MeV})$ and (8.53) $k_{\text{B}}T_{\text{RM}} = O(10\,\text{eV})$ to reach the last numerical estimate. Thus, in order to produce a $(\Omega_0 - 1) = O(1)$ now, the combined result of (9.16) and (9.17) tells us that one has to have at the epoch of primordial nuclear synthesis a density ratio equal to unity to an accuracy of one part in 10^{15}. Namely, we must have $[\Omega(t_{\text{bbn}}) - 1] = O(10^{-14})$. That the FLRW cosmology requires such an unnatural initial condition constitutes the flatness problem.

The horizon problem

Our universe is observed to be very homogeneous and isotropic. In fact, we can say that it is "too homogeneous and isotropic." Consider two different parts of the universe that are outside of each other's horizons. They are so far apart that no light signal sent from one at the beginning of the universe could have reached the other. Yet they are observed to have similar properties. This suggests their being in thermal contact sometime in the past. How can this be possible?

This horizon problem can be stated most precisely in terms of the observed isotropy of the CMB radiation (up to one part in 100,000, after subtracting out the dipole anisotropy due to the peculiar motion of our Galaxy). When pointing our instrument to measure the CMB, we obtain the same blackbody temperature in all directions. However, every two points in the sky with an angular separation on the order of a degree actually correspond to a horizon separation back at the photon-decoupling time t_γ (see (9.31)). The age of the universe at photon decoupling time was about 350,000 years, yet the observed isotropy indicates that regions far more than the horizon distance 350,000 light-year apart were strongly correlated. This is the horizon problem of the standard FLRW cosmology.

Initial conditions required for the standard cosmic evolution

We have discussed the horizon problem and flatness problem, etc. as the shortcomings of the standard big bang model. Nevertheless, it must be emphasized that they are not contradictions since we could always assume that the universe had just these conditions initially to account for the observed universe today. For example, the horizon problem can be interpreted simply as reflecting the fact that the universe must have been very uniform to begin with. These "problems" should be viewed as informing us of the correct initial conditions for the cosmic evolution after the big bang: "The initial conditions must be just so." What we need is a theory of the initial conditions. Putting it in another way, the standard big bang model is really a theory for the evolution of the universe

after the big bang. We now need a theory of the big bang **itself**. A correct theory should have the feature that it would automatically leave behind a universe with just these desired conditions.

9.2.2 The inflation scenario

The initial condition problems can be solved if, in the early moments, the universe had gone through an epoch of extraordinarily rapid expansion. This can solve the flatness problem, as any initial curvature could be stretched flat by the burst of expansion, and can solve the horizon problem if the associated expansion rate could reach superluminal speed. If the expansion rate could be greater than the light speed, then one horizon volume could have been stretched out to such a large volume that corresponded to many horizon volumes after this burst of expansion. This rapid expansion could happen if there existed then a large cosmological constant Λ, which could supply a huge repulsion to the system. The question is, then, what kind of physics can give rise to such a large Λ? In this section, we explain how modern particle physics can suggest a possible mechanism to generate, for a short instant of time, such a large vacuum-energy.

False vacuum, slow rollover phase transition and an effective Λ

The inflationary cosmology was invented in 1980 by Alan Guth in his study of the cosmological implications of the grand unified theories (GUTs) of particle interactions. The basic idea of a GUT is that particle interactions possess certain symmetry.[3] As a result, all the fundamental forces-the strong, weak, and electromagnetic interactions (except for gravity)-behave similarly at high energy. In fact they are just different aspects of the same (unified) interaction like the different faces of the same die. However, the structure of the theory is such that there is a phase transition at a temperature corresponding to the grand unification energy scale, around 10^{15}–10^{16} GeV. In the energy regime higher than this scale, the system is in a symmetric phase and the unification of particle interactions is manifest (i.e. all interactions behave similarly); when the universe cooled below this scale, the particle symmetry became hidden, showing up as distinctive forces. (For a discussion of spontaneous symmetry breakdown, that is, hidden symmetry, as illustrated by spontaneous magnetization of a ferromagnet, see Section A.3.)

In quantum field theory, particles are quantum excitations of their associated fields: electrons of the electron field, photons of the electromagnetic field, etc. New fields are postulated to exist, related to yet to be discovered particles. What brings about the above-mentioned spontaneous symmetry breaking and its associated phase transition is the existence of a certain spin-zero field $\phi(x)$, called the Higgs field, or Higgs particle. Such a field, just like the familiar electromagnetic field, carries energy. What is special about a Higgs field is that it possesses a potential energy density function $V(\phi)$ much like the potential energy function in the ferromagnet example of Section A.3. Normally one would expect field values to vanish in the vacuum state (the state with the lowest energy). A Higgs field, surprisingly, can have a nonzero vacuum state field permeating throughout in space, cf. Figs 9.3(a) and (b). The effect of this

[3] "Particle interaction symmetry" has the same meaning as "symmetry in particle physics" as explained in Chapter 1: physics equations are unchanged under some transformation. However, instead of transformations of space and time coordinates as in relativity, here one is concerned with transformation in some "internal charge space." The mathematical description of symmetry is group theory. An example of grand unification group is $SU(5)$ and particles form multiplets in this internal charge space. Members of the same multiplet can be transformed into each other: electrons into neutrinos, or into quarks, and the GUT physics equations are covariant under such transformations. After the spontaneous symmetry breaking, the interactions possess less symmetry: for example, $SU(5)$ is reduced down to $SU(3) \times SU(2) \times U(1)$, which is the symmetry group of the low energy effective theory known as the Standard Model of quantum chromodynamics and electroweak interactions.

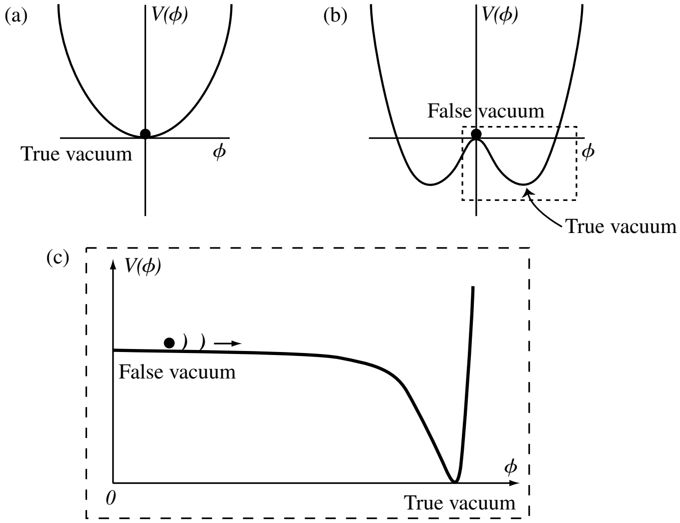

Fig. 9.3 Potential energy function of a Higgs field is illustrated by the simple case of $V(\phi) = \alpha(T)\phi^2 + \lambda\phi^4$, possessing a discrete symmetry $V(-\phi) = V(\phi)$. The parameter α has temperature-dependence, for example, $\alpha = \alpha_0(T - T_c)$, where, just like λ, α_0 is a positive constant. (a) Above the critical temperature ($T > T_c$, hence $\alpha > 0$), we have the normal case of the lowest energy state (the vacuum) being at $\phi_0 = 0$, which is symmetric under $\phi \to -\phi$. (b) Below T_c (hence $\alpha < 0$), the symmetric $V(\phi)$ has the lowest energy at points $\phi_\pm = \pm\sqrt{-\alpha/2\lambda}$ while $V(\phi = 0)$ is a local maximum. The choice of the vacuum state being either of the asymmetric ϕ_+ or ϕ_- breaks the symmetry (cf. similar plot in Fig. A.3(b)). The dashed box in (b) is displayed in (c) to show that the inflation/Higgs potential $V(\phi)$ has an almost flat portion at the $\phi = 0$ origin for a slow rollover transition. The dot represents the changing location of the system—rolling from a high plateau of the false vacuum toward the true vacuum at the bottom of the trough.

hidden symmetry can then spread to other particles through their interaction of the Higgs field. For example, a massless particle can gain its mass when propagating in the background of such a Higgs field. Different Higgs fields are posited to exist. Here we are referring to the Higgs particles in GUTs, which may have a mass $O(10^{15}\,\text{GeV}/c^2)$. These should not be confused with the electroweak Higgs particle, thought to have a mass on the order of $10^2\,\text{GeV}/c^2$, which is responsible to give masses to electrons and quarks as well as the W and Z bosons that mediate weak interactions.

In the cosmological context, such a postulated field is simply referred to as the **inflation field**, or **inflation/Higgs field**. Linde, and independently Albrecht and Steinhardt, elaborated further on the original scheme by Guth. They suggested that parameters of the unified theory were such that the potential energy function of the inflation field had a very small slope around the origin as in Fig. 9.3(c). As the universe cools, the temperature dependent parameters change so that the potential energy function changes from Fig. A.3(a) to (b). The prior lowest energy point at zero field value became a local maximum and the system would rollover to the new asymmetric vacuum state where the Higgs field would have a nonvanishing vacuum value. But the parameters are such that this rollover was slow. During this transition, we could regard the system, compared to the true (asymmetric) vacuum state, as having an extra energy density. We say the system (i.e. the universe) was temporarily in a **false vacuum**. Having this vacuum-energy density, which is time and position independent, the universe effectively had a large cosmological constant.

Exponential expansion in a vacuum-energy dominated universe

Let us consider the behavior of the scale factor $a(t)$ in a model with $\Lambda > 0$ when the matter density can be ignored. In such a vacuum-energy dominated situation, the behavior of expansion rate $\dot{a}(t)$ is such that we can always approximate the curvature signature as $k \approx 0$ (cf. (9.22)). Equation (9.8) then becomes

$$\frac{\dot{a}^2}{a^2} = \frac{8\pi G_{\mathrm{N}}}{3}\rho_\Lambda = \frac{\Lambda c^2}{3}. \tag{9.18}$$

Thus $\dot{a}$ is proportional to the scale factor a itself. Namely, we have the familiar rate equation. It can be solved to yield an exponentially expanding universe (called the **de Sitter universe**):

$$a(t_2) \equiv a(t_1)\, e^{(t_2 - t_1)/\Delta\tau} \tag{9.19}$$

with

$$\Delta\tau = \sqrt{\frac{3}{\Lambda c^2}} = \sqrt{\frac{3}{8\pi G_{\mathrm{N}}\rho_\Lambda}}, \tag{9.20}$$

where we have expressed the cosmological constant in terms of the vacuum-energy density $\rho_\Lambda c^2$ as in (9.2). Physically we can understand this exponential result because the repulsive expansion is self-reinforcing: as the energy density ρ_Λ is a constant, the more the space expands, the greater is the vacuum-energy and negative pressure, causing the space to expand even further. In fact, we can think of this Λ repulsive force as residing in the space itself, so as the universe expands, the push from this Λ energy increases as well. We note that the total energy was conserved during the inflationary epoch's rapid expansion because of the concomitant creation of gravitational field, which has a **negative** potential energy (cf. Section 8.3.1).

Remark: Because Λ represents a constant energy density, it will be the dominant factor $\rho_\Lambda \gg \rho_{\mathrm{M}}$ at later cosmic time, because the matter density ρ_{M} decreases as a^{-3}. This dominance means that it is possible for the universe to be geometrically closed ($\Omega > 1$ and $k = +1$), yet does not stop expanding. Namely, with the presence of a cosmological constant, the mass/energy density Ω (hence the geometry) no longer determines the fate of the universe in a simple way. In general, a universe with a nonvanishing Λ, regardless of its geometry, would expand forever. The only exception is when the matter density is so large that the universe starts to contract before ρ_Λ becomes the dominant term.

9.2.3 Inflation and the conditions it left behind

In the previous section we have described how the grand unification Higgs field associated with spontaneous symmetry breaking can serve as the inflation field. A patch of the universe with this "inflation/Higgs matter" might have undergone a slow rollover phase transition and thus lodged temporarily in a false vacuum with a large constant energy density. The resultant effective cosmological constant Λ_{eff} provided the gravitational repulsion to inflate the scale factor exponentially. A grand unification thermal energy scale $E_{\mathrm{GU}} = O(10^{16}\,\mathrm{GeV})$, that is, a temperature $T_{\mathrm{GU}} = O(10^{29}\,\mathrm{K})$, which according to (8.44) corresponds to the cosmic time $t_{\mathrm{GU}} \simeq 10^{-38}$ s. The energy density $\rho_{\mathrm{GU}}c^2$ can be estimated as follows: in a relativistic quantum system (such as quantum fields) there is the natural energy-length scale given by the product of Planck's constant (over 2π) times the velocity of light: $\hbar c = 1.97 \times 10^{-16}\,\text{GeV} \cdot \text{m}$. Using this conversion factor we have the energy density scale for grand unification

$$\rho_{\text{GU}} c^2 \simeq \frac{(E_{\text{GU}})^4}{(\hbar c)^3} \simeq 10^{100}\,\text{J/m}^3. \tag{9.21}$$

For a vacuum-energy density $\rho_\Lambda \approx \rho_{\text{GU}}$, the corresponding exponential expansion time-constant $\Delta\tau$ of (9.20) had the value $\Delta\tau \simeq 10^{-37}$ s. Namely, the exponential inflationary expansion took place when the universe was $t_{\text{GU}} \simeq 10^{-38}$ s old, with an exponential expansion time constant of $\Delta\tau = O(10^{-37}\,\text{s})$. By a "slow" rollover phase transition we meant that the parameters of the theory are such that inflation might have lasted much longer than 10^{-37} s, for example, 10^{-35} s (100 e-fold), expanding the scale factor by more than 30 orders of magnitude, until the system rolled down to the true vacuum, ending the inflation epoch (cf. Fig. 9.3). Afterwards the universe commenced the adiabatic expansion and cooling according to the standard FLRW model until the present epoch. This dynamics has the attractive feature that it would leave behind precisely the features that had to be postulated as the initial conditions for the standard FLRW cosmology.

The horizon and flatness problems solved

With the exponential behavior of the scale factor in (9.19), we can naturally have superluminal ($\dot{a}R_0 > c$) expansion as the rate $\dot{a}(t)$ also grows exponentially. This does not contradict special relativity, which says that an object cannot pass another one faster than c in one fixed frame. Putting it in another way, while an object cannot travel faster than the speed of light through space, there is no restriction stipulating that space itself cannot expand faster than c. Having a superluminal expansion rate, this inflationary scenario can solve the horizon problem, because two points that are a large number of horizon lengths apart now (or at the photon decoupling time when the CMB was created) could still be in causal contact before the onset of the inflationary epoch. They started out being thermalized within one horizon volume before the inflation epoch, but became separated by many horizon lengths due to the superluminal expansion.

This inflationary scenario can solve the flatness problem because the space was stretched so much that it became, after the inflationary epoch, a geometrically flat universe to a high degree of accuracy. When this exponential expansion (9.19) is applied to the Friedmann Eq. (9.15), it yields the ratio

$$\frac{1 - \Omega(t_2)}{1 - \Omega(t_1)} = \left[\frac{\dot{a}(t_2)}{\dot{a}(t_1)}\right]^{-2} = e^{-2(t_2 - t_1)/\Delta\tau}. \tag{9.22}$$

Just as the scale factor was inflated by a large ratio, say, $e^{(t_2 - t_1)/\Delta\tau} = 10^{30}$, we can have the RHS as small as 10^{-60}. Start with any reasonable value of $\Omega(t_1)$ we can still have, after the inflation, a $\Omega(t_2) = 1$ to a high accuracy. While the cosmic time evolution in the FLRW model, being determined by gravitational attraction, always enhances the curvature by driving the universe away from $\Omega = 1$ (hence the flatness problem), the accelerating expansion due to the vacuum repulsion always pushes the universe (very rapidly) toward the $\Omega = 1$ point. Thus a firm prediction by the inflationary scenario is that the universe left behind by inflation must have a flat geometry and, according to GR, a density equal to the critical value (9.15)—although it does not specify what components make up such a density.

The origin of matter/energy and structure in the universe

Besides the flatness and horizon problems, the standard FLRW cosmology requires as initial conditions that all the energy and particles of the universe be present at the very beginning. Furthermore, this hot soup of particles should have just the right amount of **initial density inhomogeneity** (density perturbation) which, through subsequent gravitational clumping, formed the cosmic structure of galaxies, clusters of galaxies, voids, etc. we observe today. One natural possibility is that such density perturbation resulted from quantum fluctuation of particle fields in a very early universe. However, it is difficult to understand how such microscopic fluctuations can bring forth the astrophysical-sized density nonuniformity required for the subsequent cosmic construction. Remarkably, the inflationary cosmology can provide us with an explanation of the origin of matter/energy, as well as the structure of the universe.

The inflation model suggests that at the beginning of the big bang a patch of the inflation/Higgs matter (smaller than the size of a proton) underwent a phase transition bringing about a huge gravitational repulsion. This is the driving force behind the space-explosion that was the big bang. While this inflation material (the Λ energy) expanded exponentially in size to encompass a space that eventually developed into our presently observed universe, its energy density remained essentially a constant. In this way more and more particle/field energy was "created" during the inflationary epoch. When it ended with the universe reaching the true vacuum, its oscillations at the trough in Fig. 9.3 showed up, according to quantum field theory, as a soup of ordinary particles. Namely, according to the inflation theory, the initial potential energy of the inflation/Higgs field (having little kinetic energy) was the origin of our universe's matter content when it was converted into relativistic particles.

The phenomenon of particle creation in an expanding universe can be qualitatively understood as follows: according to quantum field theory, the quantum fluctuations of the field system can take on the form of appearance and disappearance of particle–antiparticle pairs in the vacuum. Such energy nonconserving processes are permitted as long as they take place on a sufficiently short timescale Δt so that the uncertainty relation $\Delta E \Delta t \leq \hbar$ is not violated. In a static space, such "virtual processes" do not create real particles. However, when the space is rapidly expanding, that is, the expansion rate was larger than the annihilation rate, real particles were created.[4] Thus, inflation in conjunction with quantum field theory naturally gives rise to the phenomenon of particle creation. This hot, dense, uniform collection of particles is just the postulated initial state of the standard big bang model. Furthermore, the scale factor had increased by such a large factor that it could stretch the subatomic size fluctuation of a quantum field into astrophysical sized density perturbation to seed the subsequent cosmic structure formation. The resultant density fluctuation was random, "Gaussian," and scale-invariant, which will be discussed in Box 9.2.

Remarks: Our discussion of the inflation scenario has been couched in the language of grand unified Higgs field. It should be understood the grand unified

[4] For a related phenomenon of Hawking radiation, see Section A.2.

theories themselves have not been verified experimentally in any detail because its intrinsic energy scale of 10^{16} GeV is so much higher than the highest energy $\approx 10^3$ GeV reachable by our accelerators. On the other hand, we are confident that some version of grand unification is correct, as the simplest GUTs can already explain many puzzles of the Standard Model of particle physics, such as why strong interaction is strong, weak interaction weak, and why the quarks and leptons have the charges that they do. Nevertheless, the connection between grand unification and inflation cosmology has remained only as a suggestive possibility. It was our knowledge of the grand unification theory that allowed construction of a physically viable scenario that could give rise to an inflationary epoch. But what precisely is the inflation field, and what parameters actually govern its behavior remain as topics of theoretical discussion. The remarkable fact is that some reasonable speculation of this type can already lead to the resolution of many cosmological puzzles, and have predictions that have consistently checked with observation. As a final remark, we should also mention that it had generally been assumed that the effective cosmological constant, associated with the false vacuum, vanished at the end of the inflationary epoch. The general expectation was that the standard FLRW cosmology that followed the inflation epoch was one with no cosmological constant. Part of the rationale was that a straightforward estimate of the cosmological constant, as due to the zero-point energy of a quantum vacuum, yielded such an enormously large Λ (see Section A.4) that many had assumed that there must be some yet-to-be discovered symmetry argument that would strictly forbid a nonzero cosmological constant.

9.3 CMB anisotropy and evidence for $k = 0$

As discussed in Section 9.2.3, inflationary cosmology predicts that the spatial geometry of our universe must be flat. This prediction received more direct observational support through detailed measurement of the temperature anisotropy of the CMB radiation.

The CMB is the earliest and largest observable thing in cosmology. Its remarkable uniformity over many horizon lengths reflects its spatial origin as coming from a single pre-inflation horizon volume. Just before the photon decoupling time, the universe was a tightly bound photon–baryon fluid, and dark matter. The inflationary scenario, with its associated phenomenon of particle creation, also generated a small density perturbation on a wide range of distance scales onto this overall homogeneity. Because of gravitational instability, this nonuniform distribution of matter eventually evolved into the cosmic structure we see today. In the early universe, the gravitational clumping of baryons was resisted by photon radiation pressure. This set up an acoustic wave of compression and rarefaction with gravity being the driving force and radiation pressure the restoring force. All this took place against a background of dark matter fluctuations, which continued to grow because dark matter did not interact with radiation. Such a photon–baryon fluid can be idealized by ignoring the dynamical effects of gravitation and baryons. This leads to a sound wave speed

$$c_{\mathrm{s}} \simeq \sqrt{\frac{p}{\rho}} \simeq \frac{c}{\sqrt{3}} \tag{9.23}$$

as pressure and density being approximated by those for radiation $p \approx \rho c^2/3$. This compression and rarefaction was then translated through gravitational redshift into a temperature inhomogeneity, showing up as a series of peaks and troughs in the temperature power spectrum to be discussed in the following section.

9.3.1 Three regions of the angular power spectrum

From (8.85) and (8.88) for the correlation function, we see that the mean-square temperature anisotropy may be written for large multipole number l as

$$\left\langle \left(\frac{\delta T}{T} \right)^2 \right\rangle = \frac{1}{4\pi} \sum_{l=0}^{\infty} (2l+1) C_l \approx \int \frac{l(l+1)}{2\pi} C_l d \ln l. \tag{9.24}$$

$[(l(l+1)/2\pi)C_l]$ is approximately the power per logarithmic interval, and is the quantity presented in the conventional plot of power spectrum against a logarithmic multipole number (cf. Figs 9.4 and 9.13).

On small sections of the sky where curvature can be neglected, the spherical harmonic analysis becomes ordinary Fourier analysis in two dimensions. In this approximation the multipole number l has the interpretation as the Fourier wave number. Just as the usual Fourier wave number $k \approx \pi/x$, the multipole moment number $l \approx \pi/\theta$: large l corresponds to small angular scales with $l \approx 10^2$ corresponding to degree scale separation.

The inflationary scenario left behind density fluctuations that were Gaussian and scale invariant (cf. Box 9.2). Such an initial density perturbation, together with an assumption of a dark matter content dominated by nonrelativistic particles (the "cold dark matter" model), leads to a power spectrum as shown in Fig. 9.4. We can broadly divide it into three regions:

Region I ($l < 10^2$). This flat portion at large angular scales (the "Sachs–Wolfe plateau") corresponds to oscillations with a period larger than the age of the universe at photon decoupling time. These waves are essentially frozen in their initial configuration. The flatness of the curve reflects the scale-invariant nature of the initial density perturbation as given by the inflation cosmology (cf. Box 9.2).

Region II ($10^2 < l < 10^3$). At these smaller angular scales, smaller than the sound horizon, there had been enough time for the photon–baryon fluid to undergo oscillation. The peaks correspond to regions having higher, as well as lower, than average density. This is so because the power spectrum is the square of a_{lm} and hence indifferent to their signs. The troughs are regions with neutral compression, thus have maximum velocity (recall our knowledge of oscillators). CMB from such regions underwent a large Doppler shift. In short, it is a snapshot of the acoustic oscillations with modes (fundamental plus harmonics) having different wavelengths and different phases of oscillations. The amplitudes are related to cosmological parameters such as the baryon density Ω_B.

Region III ($l > 10^3$). Photon decoupling did not take place instantaneously, but the last scattering surface had a finite thickness. Photons diffuse out from any overdense region if it is smaller than the photon's mean free path, which

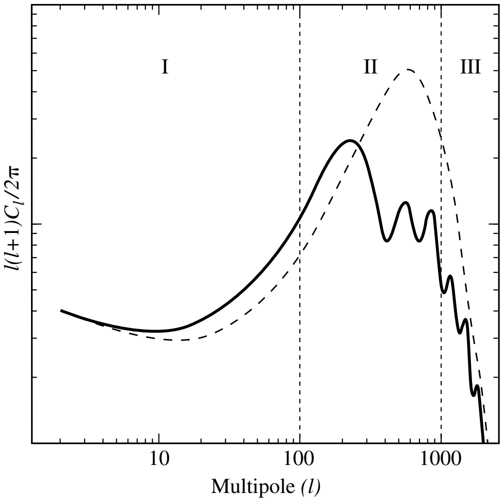

Fig. 9.4 CMB power spectrum as a function of the multipole moments. The solid curve with peaks and troughs is the prediction by inflation model (with cold dark matter). The physics corresponding to the three marked regions is discussed in the text. The dashed curve is that by the topological defect model for the origin of the cosmic structure.

was increasing as the universe expanded. The net effect was an exponential damping of the oscillation amplitude in this sub-arcminute scales.

Box 9.2 Density fluctuation from inflation is scale-invariant

Inflation produces such a huge expansion that subatomic size quantum fluctuations are stretched to astrophysical dimensions. For fluctuations larger than the sound horizon $\approx c_s H^{-1}$ one can ignore pressure gradients, as the associated sound waves cannot have crossed the perturbation in a Hubble time. The density perturbation without a pressure gradient evolves like the homogeneous universe (Problem 9.1):

$$\rho a^2(\Omega^{-1} - 1) = \text{const.}, \tag{9.25}$$

where a is the scale factor. With $\Omega \approx 1$ and $\Delta\rho \ll \rho = \rho_c \Omega$, the above relation implies

$$\rho_c a^2 \Delta\Omega = a^2 \Delta\rho = \text{const.} \tag{9.26}$$

We now consider the implication of this scaling behavior for the perturbation in gravitational potential on a physical distance scale of aL,

$$\Delta\Phi = \frac{G_N \Delta M}{aL} = \frac{4\pi}{3}\frac{G_N \Delta\rho (aL)^3}{(aL)}$$
$$= \frac{1}{2}\frac{H^2 L^2}{\rho_c} a^2 \Delta\rho,$$

which is scale invariant because of (9.26). Namely, for a physical distance scale aL, we have a gravitational potential perturbation that is independent of the scale factor a because $H = \dot{a}/a$ was a constant as both a and $\dot{a}$ undergo the same exponential increase during the inflationary epoch. Yet because the scale factor a would change by something like 30 decades during this epoch, we would have the same $\Delta\Phi$ for a range of comoving length L that changed over 30 decades. Thus, inflationary cosmology makes the strong prediction of a scale-invariant density perturbation. It can be shown that such density fluctuation, called the Harrison–Zel'dovich spectrum, would produce an angular power spectrum for the CMB anisotropy of the form

$$C_l = \frac{\text{const.}}{l(l+1)}.$$

Thus, in the plot of $l(l+1)C_l$ vs. l in Fig. 9.4 the power spectrum for the large angle region ($l < 100$) is a fairly flat curve.

In Box 9.2 above we have presented the power spectrum as predicted by the inflationary cosmology: Gaussian density perturbation leading to a random distribution of hot and cold spots on the temperature anisotropy map, and a power spectrum displaying peaks and troughs. It is illuminating to contrast this to an alternative theory of cosmic structure origin, the topological defect model. In this scenario, one posits that as the universe cooled to a thermal energy of 10^{16} GeV, the phase transition that breaks the associated grand unification symmetry also produced defects in the fabric of spacetime—in the form of strings, knots, and domain walls, etc. This introduced the initial density perturbation that seeded the subsequent structure formation. Such a density fluctuation would produce line-like discontinuities in the temperature map and a smooth power spectrum (instead of the wiggly features as predicted by the inflation model), see Fig. 9.4. As we shall discuss in the next subsection, the observed CMB anisotropy favors inflation over this topological defect model for the origin of the cosmic structure.

9.3.2 The primary peak and spatial geometry of the universe

Consider the oscillatory power spectrum in Region II of Fig. 9.4. The temperature fluctuations reflect the sound wave spectrum of the photon–baryon fluid at photon decoupling time. There would be standing waves having wavelength $\lambda_n = \lambda_1/n$, with the fundamental wavelength given by the sound horizon:

$$\lambda_1 = \int_0^{t_\gamma} \frac{c_s dt}{a(t)} \approx c_s \int_0^{t_\gamma} \frac{dt}{a(t)}. \tag{9.27}$$

Such a wavelength would appear as angular anisotropy of scale

$$\alpha_1 \simeq \lambda_1 / d(t_\gamma), \tag{9.28}$$

where $d(t_\gamma)$ is the comoving angular diameter distance from the observer to photon decoupling time. Namely, it is the comoving distance a photon would have traveled to reach us from the surface of last scattering,

$$d\left(t_\gamma\right) = c \int_{t_\gamma}^{t_0} \frac{dt}{a(t)}. \tag{9.29}$$

When evaluating the integrals in (9.27) and (9.29), we shall assume a matter-dominated flat universe with time-dependence of the scale factor $a(t) \propto t^{2/3}$

as given by (8.30),

$$\int \frac{dt}{a(t)} \propto \int a^{-1/2} da \propto a^{1/2} = (1+z)^{-1/2}. \tag{9.30}$$

Matter-domination is plausible because the radiation-matter equality time is almost an order of magnitude smaller than the photon decoupling time, that is, according to (8.68) the redshift $z_{\mathrm{RM}} \gg z_\gamma$. Thus the fundamental wavelength corresponds to an angular separation of

$$\begin{aligned}\alpha_1 &\approx \frac{\lambda_1}{d(t_\gamma)} = \frac{c_{\mathrm{s}}(1+z_\gamma)^{-1/2}}{c[(1+z_0)^{-1/2} - (1+z_\gamma)^{-1/2}]} \\ &\simeq \frac{(1+z_\gamma)^{-1/2}}{\sqrt{3}} \simeq 0.17\,\mathrm{rad} \simeq 1^\circ,\end{aligned} \tag{9.31}$$

where we have used $z_0 = 0$, $z_\gamma \simeq 1{,}100$ and, as discussed in (9.23), a sound speed $c_{\mathrm{s}} \simeq c/\sqrt{3}$. This fundamental wave angular separation in turn translates into the multipole number

$$l_1 \simeq \frac{\pi}{\alpha_1} \simeq \pi\sqrt{3}(1+z_\gamma)^{1/2} \approx 200. \tag{9.32}$$

Thus, in a flat universe we expect the first peak of the power spectrum to be located at this multipole number.

The above calculation was performed for a flat universe. What would be the result for a spatially curved universe? We will simplify our discussion by the suppression of one dimension and consider a 2D curved surface. In a positive curved closed universe ($k = +1$), light travels along longitudes (Fig. 9.5). A physical separation λ_1 at a fixed latitude, with polar angle θ and a coordinate distance $d = R_0\theta$, subtends an angle

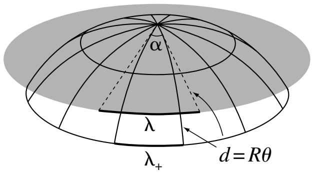

Fig. 9.5 A comparison of subtended lengths in a flat vs. positively curved surfaces. For the same angular diameter distance d, the same angle α subtends a smaller wavelength λ_+ in a closed universe when compared to the corresponding $\lambda = \lambda_+[(R_0/d)\sin(R_0/d)]^{-1} > \lambda_+$ in a flat universe.

$$\alpha_{1+} = \frac{\lambda_1}{R_0 \sin\theta} = \frac{\lambda_1}{R_0\sin(d/R_0)} = \frac{\lambda_1}{d}\left(1 + \frac{d^2}{3!R_0^2} + \cdots\right) > \frac{\lambda_1}{d}.$$

Namely, a given comoving scale (λ_1) at a fixed distance (d) the separation angle (α_{1+}) would appear to be larger (than the case of flat universe). For a negatively curved open universe ($k = -1$), one simply replaces sin by sinh:

$$\alpha_{1-} = \frac{\lambda_1}{R_0\sinh(d/R_0)} = \frac{\lambda_1}{d}\left(1 - \frac{d^2}{3!R_0^2} + \cdots\right) < \frac{\lambda_1}{d}.$$

A given comoving scale at a fixed distance, the separation angle would appear to be smaller. With the multipole number being inversely proportional to the separation angular scale, in an universe with spatial curvature the first peak would be shifted away from $l_1 \approx 200$, to a smaller (larger) multipole number for a closed (open) universe.

Although COBE satellite mapped the entire sky with high sensitivity discovering the CMB anisotropy at $\delta T/T = O(10^{-5})$, its relatively coarse angular resolution of $O(7^\circ)$ was not able to deduce the geometry of our universe. In late 1990s a number of high altitude observations, e.g., MAT/TOCO (Miller *et al.*, 1999), and balloon-borne telescopes: Boomerang (de Bernardis *et al.*, 2000), and Maxima-1 (Hanany *et al.*, 2000), had detected CMB fluctuations on smaller sizes. These observations produced evidence for a spatially flat universe by finding the characteristic size of the structure to be about a degree wide and

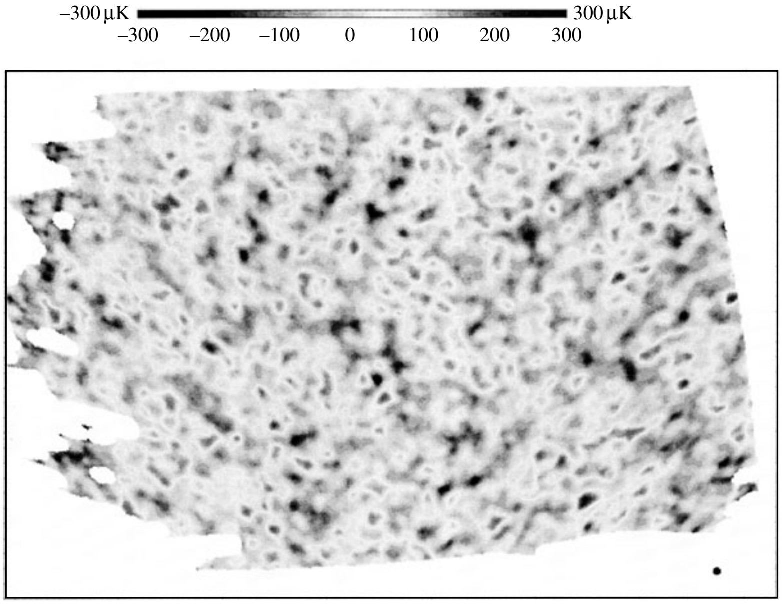

Fig. 9.6 Image of the complex temperature structure of CMB over 2.5% of the sky as captured by the Boomerang balloon-borne detector. The black dot at the lower right-hand corner represents the size of a full moon subtending an angle about half-a-degree.

a power spectrum peaked at $l \approx 200$, see Fig. 9.6. The $k = 0$ statement is of course equivalent, via the Friedmann equation, to a total density $\Omega_0 = 1$. A careful matching of the power spectrum led to

$$\Omega_0 = 1.03 \pm 0.03. \tag{9.33}$$

In the meantime, another dedicated satellite endeavor, WMAP (Wilkinson Microwave Anisotropy Probe), had reported their result in 2003. Their high resolution result allowed them to extract many important cosmological parameters: $H_0, \Omega_0, \Omega_{M,0}, \Omega_B$, and the deceleration parameter q_0, etc. (to be discussed in Section 9.5).

9.4 The accelerating universe in the present epoch

Phenomenological puzzles of a flat universe

Thus by mid/late 1990s there was definitive evidence that the geometry of the universe is flat as predicted by inflation. Nevertheless, there were several pieces of phenomenology that appeared in direct contradiction to such a picture.

A missing energy problem The Friedmann Eq. (8.7) requires a flat universe to have a mass/energy density exactly equal to the critical density, $\Omega_0 = 1$. Yet observationally, including both the luminous and dark matter, we can only find a third of this value. (Radiation energy is negligibly small in the present epoch.)

$$\Omega_M = \Omega_{LM} + \Omega_{DM} \simeq 0.30. \tag{9.34}$$

Thus, it appears that to have a flat universe we would have a “missing energy problem.”

A cosmic age problem From our discussion of the time evolution of the universe, we learned that the age of a flat universe should be two-third of the Hubble time, see (8.70),

$$(t_0)_{\text{flat}} = \frac{2}{3} t_{\text{H}} \lesssim 9\,\text{Gyr}, \tag{9.35}$$

which is shorter than the estimated age of old stars. Notably the globular clusters have been deduced to be older than 12 Gyr (cf. Section 7.1.3). Thus, it appears that to have a flat universe we would have a "cosmic age problem."

Possible resolution through a nonvanishing dark energy[5] A possible resolution of these phenomenological difficulties of a flat universe (hence inflationary cosmology) would be to assume that the cosmological constant is nonzero, even after inflation. Of course it could not have the immense size as the one it had during the inflation epoch. Rather, the constant vacuum-energy density ρ_Λ should now be about two-thirds of the critical density to provide the required missing energy.

$$\Omega = \Omega_{\text{M}} + \Omega_\Lambda \overset{?}{=} 1, \tag{9.36}$$

where $\Omega_\Lambda \equiv \rho_\Lambda/\rho_{\text{c}}$. A nonvanishing Λ would also provide the repulsion to accelerate the expansion of the universe. In an accelerating universe the expansion rate in the past must be smaller than the current rate H_0. This means that it would take a longer period to reach the present era, thus a longer age $t_0 > 2t_{\text{H}}/3$ even though the geometry is flat. This just might possibly solve the cosmic age problem mentioned as well.

9.4.1 Distant supernovae and the 1998 discovery

In order to obtain observational evidence for any changing expansion rate of the universe (i.e. to measure the curvature of the Hubble curve), one would have to measure great cosmic distances, for example, a distance method that works to over 5 billion light years. Clearly some very bright light sources are required. Since this also means that we must measure objects back in a time interval that is a significant fraction of the age of the universe, the method must be applicable to objects present at the early cosmic era. As it turns out, supernovae are ideally suited for this purpose.

SNe as standard candles

That type Ia supernovae (SNe Ia) could possibly serve as such standard candles was suggested in 1979 by Stirling Colgate. The first SN Ia was discovered in 1988 by a Danish group at redshift $z = 0.3$. At their peaks SNe Ia produce a million times more light than Cepheid Variables, the standard candle most commonly used in cosmology (cf. Section 7.3.2). SNe Ia begin as white dwarfs (collapsed old stars sustained by degenerate pressure of their electrons) with mass comparable to the sun. If the white dwarf has a large companion star, which is not uncommon, the dwarf's powerful gravitational attraction will draw matter from its companion. Its mass increases until the "Chandrasekhar limit" $\simeq 1.4\,M_\odot$. As it can no longer be countered by the electron pressure, the gravitational contraction develops and the resultant heating of the interior core would trigger the thermonuclear blast that rips it apart, resulting in an SN explosion.

[5] A dark energy is defined as the "negative equation-of-state energy", $w < 0$ in Eq. (8.4). It gives rise to a gravitational repulsion (cf. Sec. 9.1.1). The simplest example of a dark energy is Einstein's cosmological constant, with $w = -1$. NB: One should not confuse this with the energies of neutrinos, black holes, etc., which are also 'dark', but are counted as parts of the "dark matter" (cf. Sec. 7.1.4), as the associated pressure is not negative.

The supernova eventually collapses into a neutron star. Because they start with masses in a narrow range, such supernovae have comparable intrinsic brightness. Furthermore, Mark Phillips and Adam Riess and their collaborators have shown in mid-1990s that their brightness has characteristic decline from the maximum which can be used to improve on the calibration of their luminosity (the light-curve shape-analysis). Hence one has some confidence that SNe Ia can be used as standard candles. Supernovae are rare events in a galaxy. The last time a supernova explosion occurred in our galaxy was about 400 years ago. Using new technology, astronomers overcame this problem by simultaneously monitoring thousands of galaxies so that on the average some 10 to 20 supernovae can be observed in a year.

The discovery of an accelerating universe

Because light from distant galaxies was emitted long ago, to measure a star (or a supernova) farther out in distance is to probe the cosmos further back in time. An accelerating expansion means that the expansion rate was smaller in the past. Thus to reach a given redshift (i.e. recession speed) it must be located farther away[6] than expected (for a decelerating or empty universe), see Fig. 9.7. Observationally, it would be measured to be dimmer than expected.

By 1998 two collaborations: the Supernova Cosmological Project, led by Saul Perlmutter of the Lawrence Berkeley National Laboratory (Perlmutter *et al.*, 1999) and the High-z Supernova Search Team, led by Brian Schmidt of the Mount Stromlo and Siding Spring Observatory (Riess *et al.*, 1998), each had accumulated some 50 SNe Ia at high redshifts—z: 0.4–0.7 corresponding to SNe occurring five to eight billion years ago. They made the astonishing discovery that the expansion of the universe was actually accelerating, as indicated by the fact that the measured luminosities were on the average 25% less than anticipated, and the Hubble curve bent upward, Fig. 9.8.

From the Hubble curve plotted in the space of redshift and luminosity distance, one can then extract the mass and dark energy content of the universe in the present epoch. The proper distance d_p from a supernova with a redshift z in the present epoch $a(t_0) = 1$ has been shown in (7.55). Combined with the result in (7.61), this yields an expression for the luminosity distance:

$$d_\text{L}(z) = c(1+z)\int_0^z \frac{dz'}{H(z')}, \tag{9.37}$$

where, using the Friedmann Eq. (8.1), we can express the epoch-dependent Hubble constant in terms of the scale factor and the density parameters (Problem 9.2), including in particular the cosmological constant density term:

$$H(t) = H_0\left(\frac{\Omega_{\text{R},0}}{a^4} + \frac{\Omega_{\text{M},0}}{a^3} + \Omega_\Lambda + \frac{1-\Omega_0}{a^2}\right)^{1/2}, \tag{9.38}$$

where $a(t)$ can in turn be replaced by the redshift according to (7.54),

$$\begin{aligned} H(z) &= H_0[\Omega_{\text{R},0}(1+z)^4 + \Omega_{\text{M},0}(1+z)^3 + \Omega_\Lambda + (1-\Omega_0)(1+z)^2]^{1/2} \\ &\simeq H_0[\Omega_{\text{M},0}(1+z)^3 + \Omega_\Lambda + (1-\Omega_{\text{M},0}-\Omega_\Lambda)(1+z)^2]^{1/2}. \end{aligned} \tag{9.39}$$

The resultant Hubble curves $d_L(z)$ in (9.37) with $H(z)$ in the form of (9.39) that best fitted the observation data would yield values of $\Omega_{\text{M},0}$ and Ω_Λ that were

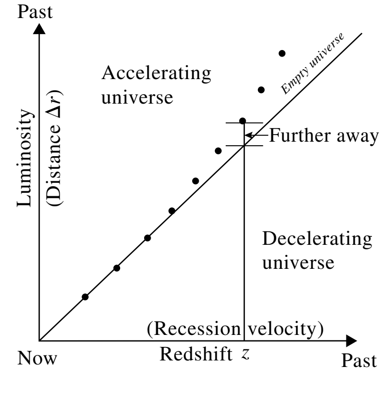

Fig. 9.7 Hubble diagram: the Hubble curve for an accelerating Universe bends upward. A supernova on this curve at a given redshift would be further out in distance than anticipated.

[6] A Hubble curve (as in Fig. 9.7) is a plot of the luminosity distance versus the redshift (measuring recession velocity). A straight Hubble curve means a cosmic expansion that is coasting. This can only happen in an empty universe (cf. Sec. 7.1.3 and Fig. 8.2). If the expansion is accelerating, the expansion rate H must be smaller in the past. From Eq. (7.5): $H\Delta r = z$, we see that, for a given redshift z, the distance Δr to the light-emitting supernova must be larger than that for an empty or decelerating universe.

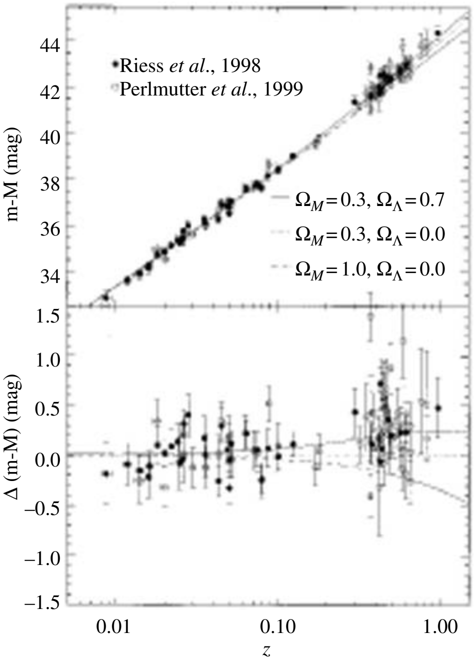

Fig. 9.8 Discovery of an accelerating universe. The Hubble plot showing the data points lying above the empty universe (dotted) line. The dashed curve represents the prediction of a flat universe without cosmological constant, the solid curve being the best fit of the observational data. The vertical axes are the luminosity distance expressed in terms of distance modulus (cf. Box 7.1). In the lower panel $\triangle(m-M)$ is the difference after subtracting out the empty universe value. Figure from review by Riess (2000).

consistent with the requirement of a flat geometry: $\Omega_{\mathrm{M},0}+\Omega_\Lambda = 1$. The favored values (see Fig. 9.9) are

$$\Omega_{\mathrm{M},0} \backsimeq 0.3 \quad \text{and} \quad \Omega_\Lambda \backsimeq 0.7 \tag{9.40}$$

suggesting that most of the energy in our universe resided in this mysterious “dark energy” (cf. sidebar 5, p. 184).

These observed values for $\Omega_{\mathrm{M},0}$ and Ω_Λ can also be translated into an age for the flat universe. Hubble constant being the rate of expansion $H = \dot{a}/a$, we can relate dt to the differential of the scale factor,

$$t_0 = \int_0^{t_0} dt = \int_0^1 \frac{da}{aH}. \tag{9.41}$$

From (9.38) for the scale-dependent Hubble constant, this yields an expression of the age[7] in terms of the density parameters:

$$t_0 = t_\mathrm{H} \int_0^1 \frac{da}{[\Omega_{\mathrm{R},0}a^{-2} + \Omega_{\mathrm{M},0}a^{-1} + \Omega_\Lambda a^2 + (1-\Omega_0)]^{1/2}}. \tag{9.42}$$

The spatially flat universe with negligible amount of radiation energy, $\Omega_0 = \Omega_{\mathrm{M},0} + \Omega_\Lambda = 1$, leads to a simple relation between cosmic time and scale factor of a given epoch:

$$\begin{aligned} t(a) &= t_\mathrm{H} \int_0^a \frac{da'}{[\Omega_{\mathrm{M},0}/a' + \Omega_\Lambda a'^2]^{1/2}} \\ &= t_\mathrm{H} \left[\frac{2}{3\sqrt{\Omega_\Lambda}} \ln \frac{\sqrt{\Omega_\Lambda a^3} + \sqrt{\Omega_{\mathrm{M},0} + \Omega_\Lambda a^3}}{\sqrt{\Omega_{\mathrm{M},0}}} \right]. \end{aligned} \tag{9.43}$$

[7] We can check the limit of (9.42) for a matter-dominated flat universe ($\Omega_{\Lambda,0} = \Omega_{\mathrm{R},0} = 0$ with $\Omega_0 = \Omega_{\mathrm{M},0} = 1$) which yields an age $t_0 = t_\mathrm{H} \int_0^1 a^{1/2} da = \frac{2}{3} t_\mathrm{H}$, in agreement with the result obtained in (8.30).

Thus for the supernovae results $\Omega_{\mathrm{M},0} \simeq 0.3$ and $\Omega_\Lambda \simeq 0.7$, we have the age of the universe

$$t_0 = t(1) \simeq 0.97 t_{\mathrm{H}} \simeq 13.2\,\mathrm{Gyr}. \tag{9.44}$$

This value clearly solves the cosmic age problem discussed on p 184.

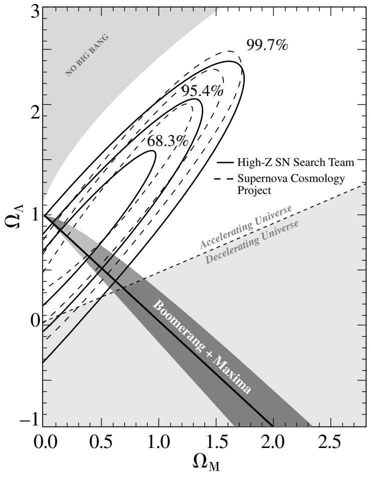

Fig. 9.9 Fitting Ω_Λ and Ω_{M} to the discovery data as obtained by High-Z SN Search Team and Supernova Cosmology Project. The favored values of $\Omega_\Lambda \simeq 0.7$ and $\Omega_{\mathrm{M}} \simeq 0.3$ follow from the central values of CMB anisotropy $\Omega_\Lambda + \Omega_{\mathrm{M}} \simeq 1$ (the straight line) and those of the SNe data represented by confidence contours (ellipses) around $\Omega_\Lambda - \Omega_{\mathrm{M}} \simeq 0.4$.

9.4.2 Transition from deceleration to acceleration

Since the immediate observational evidence from these far away supernovae is a smaller-than-anticipated luminosity, one wonders whether there is a more mundane astrophysical explanation. There may be one (or a combination of several) mundane cause that can mimic the observational effects of an accelerating universe. Maybe this luminosity diminution is brought about not because the supernovae were further away than expected, but by the absorption by yet-unknown[8] interstellar dust, and/or by some yet-unknown evolution of supernovae themselves (i.e. supernovae's intrinsic luminosity were smaller in the cosmic past)? However, all such scenarios would lead us to expect that the supernovae, at even greater distances (and even further back in time), should have their brightness continue to diminish.

For the accelerating universe, on the other hand, this diminution of luminosity would stop, and the brightness would increase at even larger distances. This is so because we expect the accelerating epoch be proceeded by a decelerating phase. The dark energy should be relatively insensitive to scale change $\rho_\Lambda \sim a^0(t)$ (the true cosmological constant is a constant density, independent of scale change), while the matter or radiation energy densities, $\rho \sim a^{-3}(t)$ or $a^{-4}(t)$, should be more and more important in earlier times. Thus, the early universe could not be dark energy dominated, and it must be decelerating. This transition from a decelerating to an accelerating phase would show up as a bulge in the Hubble curve, see Fig. 9.10.

Let us estimate the redshift when the universe made this transition. We define an epoch-dependent "deceleration parameter" which generalizes the q_0 parameter of Problem 7.11,

$$q(t) \equiv \frac{-\ddot{a}(t)}{a(t)H^2(t)},$$

which, through the Friedmann equation, can be related to the density ratios (Problem 8.10)

$$\begin{aligned} q(t) &= \Omega_{\mathrm{R}}(t) + \frac{1}{2}\Omega_{\mathrm{M}}(t) - \Omega_\Lambda \\ &= \frac{\Omega_{\mathrm{R},0}}{[a(t)]^4} + \frac{\Omega_{\mathrm{M},0}}{2[a(t)]^3} - \Omega_\Lambda. \end{aligned} \tag{9.45}$$

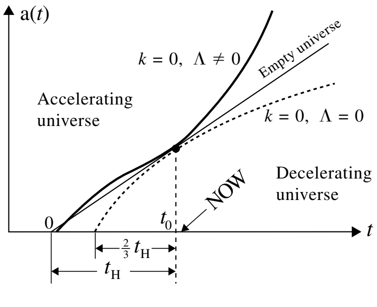

Fig. 9.10 Time evolution of an accelerating universe. It started out in a decelerating phase before taking on the form of an exponential expansion. The transition to an accelerating phase shows up as a "bulge"; this way it has an age longer than the $\Lambda = 0$ flat universe age of $\frac{2}{3}t_{\mathrm{H}}$.

After dropping the unimportant $\Omega_{\mathrm{R},0}$ and replacing the scale factor by z, we have

$$q(z) \simeq \frac{1}{2}\Omega_{\mathrm{M},0}(1+z)^3 - \Omega_\Lambda. \tag{9.46}$$

The transition from decelerating ($q > 0$) to the accelerating ($q < 0$) phase occurred at redshift $z_{\mathrm{M}\Lambda}$ (corresponding to the matter/dark-energy equality

[8] The absorption and scattering by ordinary dust shows a characteristic frequency dependence that can, in principle, be subtracted out. By the unknown dust we refer to any possible "gray dust" that could absorb light in a frequency-independent manner.

time) when the deceleration parameter vanished $q(z_{\mathrm{M}\Lambda}) \equiv 0$, or

$$1 + z_{\mathrm{M}\Lambda} = \left(\frac{2\Omega_\Lambda}{\Omega_{\mathrm{M},0}}\right)^{1/3}. \tag{9.47}$$

The supernovae data translate into a transition redshift of $z_{\mathrm{M}\Lambda} \simeq 0.7$, corresponding to a scale factor of $a_{\mathrm{M}\Lambda} \simeq 0.6$ and a cosmic time, according to (9.43), of $t_{\mathrm{M}\Lambda} = t(a = 0.6) \simeq 7\,\mathrm{Gyr}$—in cosmic terms, the transition took place only recently ("just yesterday")! This reflects the fact that the matter density in the present epoch $\Omega_{\mathrm{M},0}$ happens to be comparable to the dark energy density Ω_Λ.

Thus, the conclusive evidence for the accelerating universe interpretation of the supernovae data is observed in this bulge structure, which cannot be mimicked by any known astrophysical causes. The 1998 discovery data (z: 0.4–0.7) showed the rise of this bulge, but we need to see the falling part of the Hubble curve. SNe further out ($z > 0.7$) should be still in the decelerating phase; they should be brighter than what is expected of continuing dimming scenario that mundane interpretation would have us anticipate. Astonishingly, just such an early decelerating phase had been detected.

After the original discovery of an accelerating universe, researchers had searched for other supernovae at high z. The supernova labeled SN1997ff had been serendipitously recorded by the Hubble Space Telescope, and by other observational means (some intentionally, and some unpremeditated). Through a major effort at data analysis, its properties were deduced in 2001, showing that it is a type Ia SN having a redshift of $z \simeq 1.7$ and, thus an explosion occurring 10 billion years ago, making it by far most distant supernova ever detected. Remarkably, it is brighter by almost a factor of two compared to the expectation of continual dimming, see Fig. 9.11. This is the bulge feature unique to a Hubble curve for an accelerating universe—the light was emitted long ago, when the expansion of the universe was still decelerating. During the 2001–03 period, many more high-z SNe had been discovered both from ground-based observation and with Hubble Space Telescope. These data had provided conclusive evidence for cosmic deceleration that preceded the present epoch of cosmic acceleration (Riess *et al.*, 2004).

The problem of interpreting Λ as quantum vacuum-energy The introduction of the cosmological constant in the GR field equation does not explain its physical origin. In the inflation model the effective cosmological term represents the false vacuum-energy of an inflation field. In fact, the cosmological constant

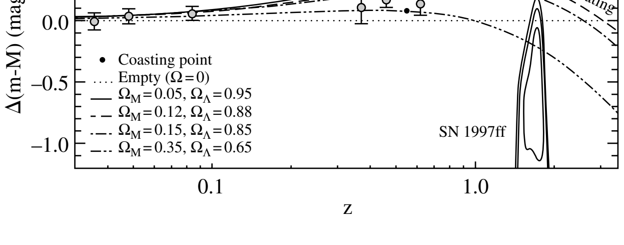

Fig. 9.11 Discovery of the decelerating phase. Graph from (Riess *et al.*, 2001). Location of SN1997ff (because of measurement uncertainties, shown as a patch on the right side of the diagram) and other high z SNe are plotted with respect to those for an empty universe (the horizontal line) in a Hubble diagram (cf. lower panel of Fig. 9.8). The black spots follow an up-turning curve which represents the luminosity and redshift relation showing continuing dimming as a mundane astrophysical explanation would require.

has a more fundamental physical interpretation—as the quantum mechanical vacuum-energy[9] (also called the zero-point energy). From the view of quantum field theory, a vacuum state is not simply "nothingness." The uncertainty principle informs us that the vacuum has a constant energy density.[10] However, as we show in Section A.4 such an association leads to an estimate of ρ_Λ that is something like 10^{120} larger than the observed value ($\rho_\Lambda \simeq \rho_c$, the critical density). Since it is off-the-mark by such a large factor (there are very few numbers in physics as large as 10^{120}), many thought that there must be some yet-undiscovered-symmetry principle which would demand the quantum vacuum-energy to be exactly zero.[11] The dark energy driving the accelerating expansion (cf. sidebar 5, p. 184) is suggested to have its physical origin in something other than quantum zero-point energy. One possibility is that the dark energy (with a density parameter $\Omega_X \simeq 0.7$) is associated with some yet-unknown scalar field (sometimes referred to as the "quintessence"), somewhat akin to the association of the inflationary expansion to the inflation/Higgs field. Such theories often have an equation-of-state parameter $w_X \neq w_\Lambda = -1$. However, observational data do not support a dark energy w_X significantly different from the value of -1. For example, the deceleration parameter q_0 can be independently measured. Then the relation in Problem 8.10, from which (9.45) was derived, implies

$$q_0 = \frac{1}{2}[\Omega_{M,0} + (1 + 3w_X)\Omega_X] \tag{9.48}$$

or

$$w_X = \frac{1}{3}\left(\frac{2q_0 - \Omega_{M,0}}{\Omega_X} - 1\right). \tag{9.49}$$

Thus $\Omega_X \simeq 0.7$, $\Omega_{M,0} \simeq 0.3$, and an observed value of $q_0 \simeq -0.6$ (see Table 9.1) would lead to $w_X = -0.95$. The current observational data are certainly consistent with the dark energy having just the property of the cosmological constant as first theorized by Einstein.

9.5 The concordant picture

An overall coherent and self-consistent picture of the cosmos has emerged that can account for the geometry and structure of the universe, as well as its evolution onward from a fraction of a second after the big bang. In this section, we first summarize the cosmological parameters and discuss the concordant cosmological model that had emerged. Even though we have a consistent picture, there are still many unsolved problems; we shall mention some of them at the end of this chapter.

Ten cosmological parameters

Our previous discussion has concentrated on conceptually and technically simpler approaches in obtaining cosmological parameters—counting and weighing methods, plotting the Hubble curve (including data from high-redshift supernovae), and light-element abundance, etc. These measurements have now been confirmed and hugely improved by the analysis of very different physical phenomena: the CMB temperature anisotropy (in particular as measured by WMAP) in combination with analysis of large-scale structure survey data

[9] The inflationary cosmology discussion presupposes that the quantum vacuum contribution to the cosmological constant is negligibly small.

[10] In fact, QFT also pictures the vacuum as a sea of sizzling activities with constant creation and annihilation of particles.

[11] In Section A.4 we also illustrate this by the example of "supersymmetry," the invariance between half-integer spin particles (fermions) and integer spin particles (bosons).

(obtained in particular by 2dF and SDSS). While a presentation of the analysis involved in the large-scale structure study is beyond the scope of this book, we have briefly discussed the CMB anisotropy (cf. Sections 8.5.4 and 9.3): detailed study of the power spectrum through a spherical harmonics decomposition can be displayed as a curve (relative amplitude vs. angular momentum number) with a series of peaks. The primary peak (i.e. the dominant structure) is at the one degree scale showing that the spatial geometry is flat; the secondary peaks are sensitive to other cosmological parameters such as the baryon contents of the universe, $\Omega_B \simeq 0.04$, etc. WMAP has a much improved angular resolution compared to COBE, Fig. 9.12, this allowed us to add an array of cosmological parameters (Table 9.1).

The standard model of cosmology

Cosmology has seen a set of major achievements over the past decade, to the extent that something like a standard model for the origin and development of the universe is now in place: the FLRW cosmology proceeded by an inflationary epoch. Many of the basic cosmological parameters have been deduced in several independent ways, arriving at a consistent set of results. These data are compatible with our universe being infinite and spatially flat, having matter/energy density equal to the critical density, $\Omega_0 = 1$. The largest energy component $\Omega_X \simeq 0.7$ is consistent with it being Einstein's cosmological constant $\Omega_X = \Omega_\Lambda$. In the present epoch this dark energy content is comparable in size to the matter density $\Omega_M \simeq 0.3$, which is made up mostly of cold dark matter. The expansion of the universe will never stop—in fact having entered the accelerating phase, the expansion will be getting faster and faster.

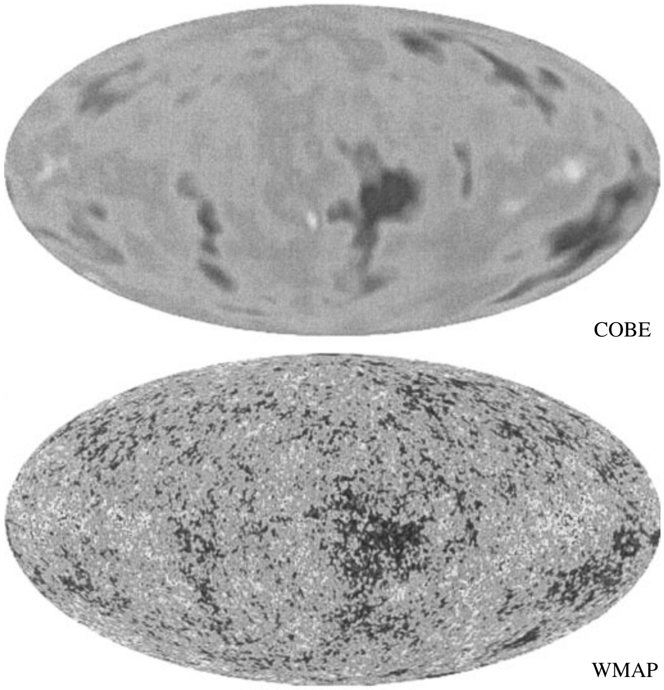

Fig. 9.12 The temperature fluctuation of CMB is a snap-shot of the baby universe at photon decoupling time. A comparison of the results by COBE vs. WMAP shows the marked improvement in resolution by WMAP. This allowed us to extract many more cosmological parameters from the latest observations.

Table 9.1 Ten cosmological parameters (from Freedman and Turner 2003). The left column is the combined analysis of published data; the right column the first-year data from WMAP (Bennett *et al.*, 2003). The equation numbers in the central column refer to part of the text, where such parameters were discussed. The first parameter h_0 is the Hubble constant H_0 measured in units of 100 (km/s)/Mpc

	Parameter value	Description	WMAP
h_0	0.72 ± 0.07	Present expansion rate (7.7)	$0.71^{+0.04}_{-0.03}$
q_0	-0.67 ± 0.25	Deceleration parameter (9.48)	-0.66 ± 0.10
t_0	13 ± 1.5 Gyr	Age of the universe (9.44)	13.7 ± 0.2 Gyr
T_0	2.725 ± 0.001 K	CMB temperature (8.64)	
Ω_0	1.03 ± 0.03	Density parameter (9.33)	1.02 ± 0.02
Ω_B	0.039 ± 0.008	Baryon density (8.58)	0.044 ± 0.004
Ω_{CDM}	0.29 ± 0.04	Cold dark matter density (7.25)	0.23 ± 0.04
Ω_ν	$0.001 - 0.05$	Massive neutrino density	
Ω_X	0.67 ± 0.06	Dark energy density (9.40)	0.73 ± 0.04
w_X	-1 ± 0.2	Dark energy equation of state (9.49)	< -0.8

Still many unsolved problems

Although we have a self-consistent cosmological description, many mysteries remain. We do not really know what makes up the bulk of the dark matter, even though there are plausible candidates as predicted by some yet-to-be-proven particle physics theories. The most important energy component is the mysterious "dark energy," although a natural candidate is the quantum vacuum-energy. Such an identification leads to an estimate of its size that is completely off the mark (cf. Section A.4). If one can show that the quantum vacuum-energy must somehow vanish due to some yet-to-be-found symmetry principle, a particular pressing problem is to find out whether this dark energy is time-independent, as is the case of the cosmological constant, or is it more like an effective Lambda coming from some quintessence scalar field like the case of inflation. Despite our lack of understanding of this dark energy, the recent discoveries constitute a remarkable affirmation of the inflationary theory of the big bang. Still, even here the question remains as to the true identity of the inflation/Higgs field. We need to find ways to test the existence of such a field in some noncosmological settings.

Besides the basic mystery of dark energy ("the cosmological constant problem") there are other associated puzzles, one of them being the "cosmic coincidence problem": we have the observational result that in the present epoch the dark energy density is comparable to the matter density, $\Omega_X \backsimeq \Omega_M$. Since they scale so differently ($\Omega_M \sim a^{-3}$ vs. $\Omega_X \sim a^0$) we have $\Omega_M \simeq 1$ in the cosmic past, and $\Omega_\Lambda \simeq 1$ in the future. Thus, the present epoch is very special—the only period when they are comparable. Then the question is why? How do we understand this requirement of fine tuning the initial values in order to have $\Omega_M \backsimeq \Omega_X$ now?

A finite dodecahedral universe: a cautionary tale

It cannot be emphasized too much that the recent spectacular advances in cosmology have their foundation in the ever-increasing amount of high precision observational data. Ultimately any cosmological theory will stand or

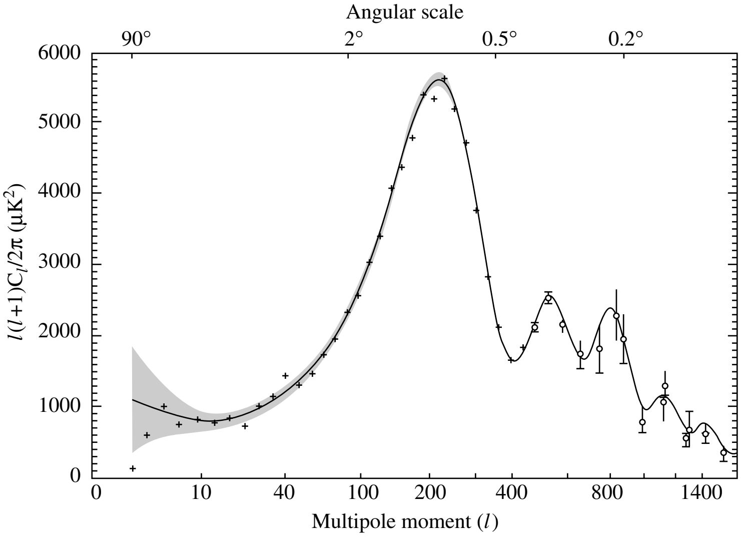

Fig. 9.13 The angular power spectrum of CMB temperature anisotropy. The dots are the first-year data-points from WMAP. The theoretical curve follows from inflationary model (having cold dark matter) with parameters given in Table 9.1. The fan-shaped shaded area at low multiple moments reflects the uncertainty due to cosmic variance, cf. (8.89).

fall, depending on its success in confronting experimental data. In this context we offer the following cautionary tale.

An inspection of the CMB power spectrum in Fig. 9.13 shows that a few data points in the large angle (low l) region tend to be lower than the theoretical curve based on the standard cosmological model outlined above. This does not concern most cosmologists because they are still in the shaded area corresponding to the statistical uncertainty called cosmic variance (cf. (8.89)). Nevertheless, it is possible to interpret these low data points as potential signature of a finite universe. The weakness of quadrupole ($l = 2$) and octupole term ($l = 3$) can be taken as lack of temperature correlation on scales greater than 60°. Maybe the space is not infinite and the broadest waves are missing because space is not big enough to accommodate them. Our discussion above has shown the evidence for the space being locally homogenous and isotropic. However, local geometry constrains, but does not dictate, the shape of the space. Thus, it is possible that the topology of the universe is nontrivial. Luminet *et al.* (2003) constructed just such a model universe based on a finite space with a nontrivial topology (the Poincaré dodecahedral space). It has a positive curvature with $\Omega_0 = 1.013$, which is compatible with observation as listed in Table 9.1. One of the ways to study the shape, or topology, of the universe is based on the idea that if the universe is finite, light from a distant source will be able to reach us along more than one path. This will produce matching images (e.g. circles) in the CMB anisotropy. A search for such matching circles has turned out to be negative (Cornish *et al.*, 2004). Thus this finite universe model may in the end, be ruled out by observation.

Our purpose in reporting this particular episode in the cosmological study is to remind ourselves of the importance of keeping an open mind of alternative cosmologies. This example showed vividly how drastically different cosmological pictures can be based on cosmological parameters that are not that different from each other. Thus, when looking at a result such as $\Omega_0 = 1.03 \pm 0.03$ we should refrain from jumping to the conclusion that data has already shown a $\Omega_0 = 1$ flat universe. This shows the importance of acquiring high precision data, which will ultimately decide which model gives us the true cosmology. On the other hand, while slight change of one or two parameters may favor different cosmological models, it is the overall theoretical consistency and the ability to account for a whole array of data in cosmology that ultimately allows us to believe that the current concordant picture has a good chance to survive future experimental tests.

Review questions

1. Use the first law of thermodynamics to show the constancy of a system's energy density (even as its volume changes) requires this density to be equal to the negative of its pressure.
2. A vacuum-energy dominated system obey Newton's equation $\nabla^2\Phi = -\Lambda c^2$, where Λ is a positive constant. What is the gravitational potential $\Phi(r)$ satisfying this equation? From this, find the corresponding gravitational field $\mathbf{g}(r) \equiv -\nabla\Phi(r)$.
3. From the Friedmann Equation $[1 - \Omega(t)] = -kc^2/[\dot{a}(t)R_0]^2$ and the fact that the universe has been matter-dominated since the radiation–matter equality time with redshift $z_{\text{RM}} = O(10^4)$, show that the deviation of energy density ratio Ω from unity at t_{RM} must be a factor of 10,000 times smaller than that at the present epoch t_0:

   $$[1 - \Omega(t_{\text{RM}})] = [1 - \Omega(t_0)] \times 10^{-4}.$$

   Use this result (and its generalization) to explain the flatness problem.
4. What is the horizon problem? Use the result that the angular separation corresponding to one horizon length at the photon decoupling time is about one degree (for a flat universe) to explain this problem.
5. Use a potential energy function diagram to explain the idea of a phase transition in which the system is temporarily in a "false vacuum." How can such a mechanism be used to give rise to an effective cosmological constant?
6. Give a simple physical justification of the rate equation obeyed by the scale factor $\dot{a}(t) \propto a(t)$ in a vacuum-energy dominated universe. Explain how the solution $a(t)$ of such a rate equation can explain the flatness and horizon problems.
7. How does the inflationary cosmology explain the origin of mass and energy in the universe as well as the origin of the cosmic structure we see today?
8. The CMB power spectrum can be divided into three regions. What physics corresponds to each region?
9. How can the observed temperature anisotropy of the CMB be used to deduce that the average geometry of the universe is flat?
10. The age of a flat universe without the cosmological constant is estimated to be $\frac{2}{3}t_{\text{H}} \approx 9\,\text{Gyr}$. Why can an accelerating universe increase this value?
11. What is a dark energy? How is it different from the dark matter?
12. Give two reasons to explain why type Ia supernovae are ideal "standard candles" for large cosmic scale measurements.
13. Why should the accelerating universe lead us to observe the galaxies, at a given redshift, to be dimmer than expected (in an empty or decelerating universe)?
14. Why is the observation of supernovae with the highest redshifts (>0.7) in the decelerating phase taken to be the convincing evidence that the accelerating universe interpretation of SNe data (z: 0.2–0.7) is correct?
15. What is the cosmic coincidence problem?
16. What is the "standard model of cosmology"? In this model is the space finite or infinite? What is its geometry? How old is the universe? What is the energy/matter content of the universe?

Problems

(9.1) **Another form of the expansion equation** Use either the Friedmann equation or its quasi-Newtonian analog of (8.11) to derive (9.25).

(9.2) **The epoch-dependent Hubble constant and $a(t)$** Use (8.7) to replace the curvature parameter k in the Friedmann Equation (8.1) to show the epoch dependence of the Hubble constant through its relation to the density parameters as in (9.38).

(9.3) **Luminosity distance and redshift in a flat universe** Knowing the redshift-dependence of the Hubble constant from Problem 9.2 in a flat universe with negligible $\Omega_{R,0}$, show that the Hubble curve $d_L(z)$ can be used to extract the density parameters Ω_M and Ω_Λ from the simple relation

$$d_L(z) = c(1+z)\int_0^z \frac{cdz'}{H_0[\Omega_{M,0}(1+z')^3+\Omega_\Lambda]^{1/2}}.$$

(9.4) **Negative Λ and the "big crunch"** Our universe is spatially flat with the dominant component being matter and positive dark energy. Its fate is an unending exponential expansion. Now consider the same flat universe but with a negative dark energy $\Omega_\Lambda = 1 - \Omega_{M,0} < 0$, which provides a gravitational attraction [cf. (9.7)]. Show that this will slow the expansion down to a standstill when the scale factor reaches $a_{\max} = (-\Omega_\Lambda/\Omega_{M,0})^{1/3}$. The subsequent contraction will reach the big crunch $a(t_*) = 0$ at the cosmic time $t_* = \frac{2}{3}\pi t_H(-\Omega_\Lambda)^{-1/2}$.

(9.5) **Another estimate of deceleration/acceleration transition time** Another simple way to estimate the deceleration/acceleration transition ("inflection") time as the epoch when the matter and dark energy components are equal. Show that the redshift result obtained in this way is comparable to that of (9.47).

RELATIVITY
Full Tensor Formulation

Part
III

This page intentionally left blank

Tensors in special relativity

10

- We introduce the mathematical subject of tensors in a general coordinate system and apply it to the four-dimensional continuum of Minkowski spacetime.
- When physics equations are written as 4-tensor equations, they are automatically unchanged under coordinate transformation, and hence respect the principle of relativity. Such formalism is said to be "manifestly covariant."
- In the case of special relativity (SR), the coordinate transformations are Lorentz transformations.
- x^μ, ∂_μ, U^μ, and p^μ are the displacement, the gradient, the velocity, and the momentum 4-vectors; the six components of the electric and magnetic fields E_i and B_i form an antisymmetric tensor $F_{\mu\nu} = -F_{\nu\mu}$.
- The Maxwell equation and charge conservation equation are presented in manifestly covariant form.
- The energy–momentum tensor of a field system $T_{\mu\nu} = T_{\nu\mu}$ is introduced and the physical meaning of its components discussed.

In the introductory Chapter 1 we emphasized the approach of relativity as the coordinate symmetry. The principle of relativity says that physics equations should be covariant under coordinate transformations. To ensure that this principle is **automatically** satisfied, all one needs to do is to write physics equations in terms of tensors. Tensors are mathematical objects having definite transformation properties under coordinate transformations. The simplest examples are scalars and vectors. If every term in an equation has the same tensor property, that is, transforms in the same way under coordinate transformations, then the relational form of the equation will not be altered under such transformations. In the next three chapters, the full tensor formalism, hence the symmetry viewpoint of relativity, will be explicated. In this chapter, we deal mainly with basic tensors in the 4-dimensional (4D) spacetime. The formalism will be adequate for global Lorentz transformations which are relevant for special relativity (SR). In the next chapter, we discuss the topic of tensor equations that are covariant under the local (position-dependent) transformation of general relativity.

10.1 General coordinate systems

Referring back to Section 2.3.1 where we first introduced general coordinates, we recall that, in contrast to the Cartesian coordinate system in the Euclidean space, generally **coordinate basis vectors** may not be mutually perpendicular and may have different length: $\mathbf{e}_i \cdot \mathbf{e}_j \equiv g_{ij} \neq \delta_{ij}$. This means that the bases $\{\mathbf{e}_i\}$, as well as the metric g_{ij}, are not their own inverse. We have inverse bases and inverse metric, distinctive from the bases and their metric. Let us denote the **inverse bases** by a set of vectors $\{\mathbf{e}^i\}$. Our notation is that the bases have subscript labels, while inverse bases have superscript labels. The inverse relation is expressed as an orthonormality condition through dot products much like (2.33):

$$\mathbf{e}_i \cdot \mathbf{e}^j = \delta_i^j. \tag{10.1}$$

Furthermore, we have the completeness condition of $\sum_i \mathbf{e}_i \otimes \mathbf{e}^i = \mathbf{1}$, where the symbol ⊗ stands for "direct product." There are n basis vectors, $\{\mathbf{e}_i\}$ with $i = 1, 2, \ldots, n$. Each of the basis vectors in turn has components $(e_i)_a$ with $a = 1, 2, \ldots, n$. Equation (10.1) can be written out as $\sum_a (e_i)_a (e^j)_a = \delta_i^j$, while the completeness condition corresponds to the component multiplication of

$$\sum_i (e_i)_a (e^i)_b = \delta_{ab}. \tag{10.2}$$

(See Problem 10.1 for an illustrative example.) The dot products of the bases are the metric functions:

$$\begin{aligned} \mathbf{e}_i \cdot \mathbf{e}_j &\equiv g_{ij} \qquad \text{metric}, \\ \mathbf{e}^i \cdot \mathbf{e}^j &\equiv g^{ij} \qquad \text{inverse metric}. \end{aligned} \tag{10.3}$$

These metric matrices are inverse to each other,

$$g_{ik} g^{kj} = \delta_i^j \tag{10.4}$$

through the condition (10.1). Again, we shall follow the **Einstein summation convention** of summing over repeated upper and lower indices.

Covariant and contravariant vectors

Because there are two sets of coordinate basis vectors $\{\mathbf{e}_i\}$ and $\{\mathbf{e}^i\}$, for each vector $\mathbf{V}$ there can be two possible expansions:

Expansion of **V**	Projections	Component names
$\mathbf{V} = V^i \mathbf{e}_i$	$V^i = \mathbf{V} \cdot \mathbf{e}^i$	**Contravariant** components of **V**
$\mathbf{V} = V_i \mathbf{e}^i$	$V_i = \mathbf{V} \cdot \mathbf{e}_i$	**Covariant** components of **V**

(10.5)

Repeated indices are summed in the expansions of the vector **V**. We shall often refer to the contravariant and covariant components of a vector, for simplicity, as contravariant vector and covariant vector. For a general rectilinear coordinate system in a flat plane (Fig. 10.1), these two types of components can be visualized as follows: the contravariant components are the parallel projections of a vector onto the basis vectors, while the covariant components are the perpendicular projections (hence parallel projections with respect to the inverse bases).

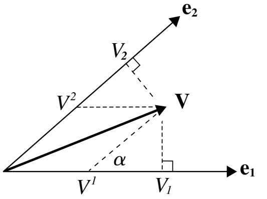

Fig. 10.1 Contravariant components (V^1, V^2) and covariant components (V_1, V_2) of a vector **V** in a general coordinate system. For the simple case of basis vectors having unit length, they are seen to be related as $V_1 = (V^1 + V^2 \cos\alpha)$ and $V_2 = (V^1 \cos\alpha + V^2)$.

One of the principal advantages for introducing these two types of tensor components is the simplicity of the resultant scalar product, which, after using (10.1), can always be expressed as

$$\mathbf{V} \cdot \mathbf{U} = (V^i \mathbf{e}_i) \cdot (U_j \mathbf{e}^j) = V^i U_j (\mathbf{e}_i \cdot \mathbf{e}^j) = V^i U_i. \tag{10.6}$$

If we had used expansion only in terms of the basis $\{\mathbf{e}_i\}$, or only the inverse basis $\{\mathbf{e}^i\}$, we would have to display the metric tensors:

$$\mathbf{V} \cdot \mathbf{U} = g_{ij} V^i U^j = g^{ij} V_i U_j. \tag{10.7}$$

A comparison of (10.6) and (10.7) shows that tensor indices can be raised or lowered through contractions with the metric tensors:

$$V_i = g_{ij} V^j, \quad V^i = g^{ij} V_j. \tag{10.8}$$

As have already used in previous chapters (e.g. (4.8)), (10.7) can be taken as our definition of the metric, especially the vectors are taken to be infinitesimal displacement vectors $d\mathbf{x} \cdot d\mathbf{x} = ds^2 = g_{ij} dx^i dx^j$.

Coordinate transformations

Under a coordinate change, the transformation of contravariant components may be written as

$$\begin{pmatrix} V^1 \\ V^2 \\ \vdots \\ V^n \end{pmatrix} \longrightarrow \begin{pmatrix} V'^1 \\ V'^2 \\ \vdots \\ V'^n \end{pmatrix} = \begin{pmatrix} L^1_1 & L^1_2 & \dots & L^1_n \\ L^2_1 & L^2_2 & & \\ \vdots & & & \\ L^n_1 & \dots & & L^n_n \end{pmatrix} \begin{pmatrix} V^1 \\ V^2 \\ \vdots \\ V^n \end{pmatrix}.$$

V^is represent the components of the vector $\mathbf{V}$ in the original coordinate system, while V'^is are those with respect to the transformed system. This relation can be written in a more compact notation as (cf. Box 2.1 and Section 2.3.1)

$$V^i \longrightarrow V'^i = [\mathbf{L}]^i_j \, V^j. \tag{10.9}$$

It is important to keep in mind that contravariant and covariant components transform differently under a coordinate transformation. We will represent the transformation of covariant components as

$$V_i \longrightarrow V'_i = [\bar{\mathbf{L}}]^j_i V_j \tag{10.10}$$

where $[\bar{\mathbf{L}}]$ in (10.10) differs from $[\mathbf{L}]$ in (10.9). The transformation property of a general tensor component can be illustrated by an example

$$T^k_{ij} \longrightarrow T'^k_{ij} = [\bar{\mathbf{L}}]^l_i \, [\bar{\mathbf{L}}]^m_j \, [\mathbf{L}]^k_n \, T^n_{lm}. \tag{10.11}$$

Namely, for each superscript index we have an $[\mathbf{L}]$ factor, and each subscript index an $[\bar{\mathbf{L}}]$. Only if T^k_{ij} transform in this way, can they be considered as tensor components. For example, the inverse metric g^{ij} having two contravariant indices must transform as

$$g'^{ij} = [\mathbf{L}]^i_k [\mathbf{L}]^j_l g^{kl}, \tag{10.12}$$

while the metric tensor transforms as

$$g'_{ij} = [\bar{\mathbf{L}}]^k_i [\bar{\mathbf{L}}]^l_j g_{kl}. \tag{10.13}$$

The $[\mathbf{L}]$ and $[\bar{\mathbf{L}}]$ transformations are in fact inverse to each other. We can see this from the invariance of the scalar product $\mathbf{U} \cdot \mathbf{V}$:

$$U'_k V'^k = U_i V^i \quad \text{or} \quad U_j [\bar{\mathbf{L}}]^j_k [\mathbf{L}]^k_i V^i = U_j \delta^j_i V^i. \tag{10.14}$$

Thus

$$[\bar{\mathbf{L}}]^j_k [\mathbf{L}]^k_i = \delta^j_i. \tag{10.15}$$

Written as matrix relations, (10.15) is just $[\bar{\mathbf{L}}][\mathbf{L}] = \mathbf{1}$. Equation (10.13) shows that the transformation of the metric involves the inverse transformation matrices as first shown in (2.45), as well as in Problem 4.3.

Remark: In a flat space, such as the Minkowski space of SR, for which we can always use a coordinate system so that the metric is position independent, we also have an invariant $\mathbf{g}' = \mathbf{g}$. Equation (10.13) becomes $\mathbf{g} = [\bar{\mathbf{L}}]\mathbf{g}[\bar{\mathbf{L}}^\top]$ which was used in Section 2.3.2 to derive the explicit form of Lorentz transformation. We should emphasize that the condition $g'^{ij} = g^{ij}$ is not possible for a curved space. While the tensor formalism remains basically the same, there is the key difference of basis vectors being necessarily position-dependent, $\{\mathbf{e}_i(x)\}$. Consequently, the metric tensors $g_{ij} = \mathbf{e}_i \cdot \mathbf{e}_j$ must also be position-dependent, and they always transform nontrivially under coordinate transformations, $g'_{ij} \neq g_{ij}$.

10.2 Four-vectors in Minkowski spacetime

As discussed in Section 2.3, Lorentz transformations may be viewed as "rotations" in the 4D spacetime. Hence if physical quantities are represented by 4-vectors and 4-tensors, the resultant physics equations will automatically be invariant under Lorentz transformation—these equations will be, manifestly, relativistic. The position-time components x^μ (the Greek index μ ranges from 0 to 3) are naturally contravariant components of the 4-position vector because they are the coefficients of expansion of the position-time vector onto the basis axes,[1] $\mathbf{x} = x^\mu \mathbf{e}_\mu$:

$$x^\mu = (x^0, x^1, x^2, x^3) = (ct, x, y, z). \tag{10.16}$$

For Minkowski spacetime the space–time interval

$$s^2 = -(x^0)^2 + (x^1)^2 + (x^2)^2 + (x^3)^2 = \eta_{\mu\nu} x^\mu x^\nu \tag{10.17}$$

has the same value with respect to every inertial observer. This can be interpreted as defining the metric for the flat Minkowski space,

$$g_{\mu\nu} = \text{diag}(-1, 1, 1, 1) \equiv \eta_{\mu\nu}. \tag{10.18}$$

[1] For the 4D space $\mu = 0, 1, 2, 3$ the frame formed by the set of four basis vectors $\{\mathbf{e}_\mu\}$ is referred to as a **tetrad** (or, in German, as a **Vierbein**).

Under a Lorentz transformation, the new components are related to the original ones by

$$x^\mu \rightarrow x'^\mu = [\mathbf{L}]^\mu_\nu x^\nu. \tag{10.19}$$

Specifically, for a boost with a velocity v in the $+x$ direction, we have (cf. (2.60))

$$\begin{pmatrix} ct' \\ x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} \gamma & -\beta\gamma & & \\ -\beta\gamma & \gamma & & \\ & & 1 & \\ & & & 1 \end{pmatrix} \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix}, \tag{10.20}$$

where β and γ are defined in (2.14). Because the metric is given by (10.18) the covariant displacement vector (in contrast to contravariant vector) is given by:

$$x_\mu = \eta_{\mu\nu} x^\nu = (-ct, x, y, z). \tag{10.21}$$

The contraction between contravariant and covariant components is the invariant interval and is related to the proper time τ:

$$s^2 = x^\mu x_\mu = -c^2t^2 + x^2 + y^2 + z^2 = -c^2\tau^2. \tag{10.22}$$

Now let's consider the coordinate transformation $[\bar{\mathbf{L}}]$ for the covariant components,

$$V_\mu \longrightarrow V'_\mu = [\bar{\mathbf{L}}]^\nu_\mu V_\nu. \tag{10.23}$$

$[\mathbf{L}]$ and $[\bar{\mathbf{L}}]$ are inverse to each other. For a boost transformation $[\mathbf{L}]$ as given in (10.20) the corresponding $[\bar{\mathbf{L}}]$ transformation can be obtained by the replacement of $(\beta \rightarrow -\beta)$:

$$[\bar{\mathbf{L}}] = \begin{pmatrix} \gamma & \beta\gamma & & \\ \beta\gamma & \gamma & & \\ & & 1 & \\ & & & 1 \end{pmatrix}. \tag{10.24}$$

The del operator

From calculations in Chapter 2 (see Problem 2.3 in particular) we know that the 4-gradient operator transforms according to an inverse Lorentz transformation

$$\frac{\partial}{\partial x^\mu} \longrightarrow \frac{\partial}{\partial x'^\mu} = [\bar{\mathbf{L}}]^\nu_\mu \frac{\partial}{\partial x^\nu}. \tag{10.25}$$

Equation (10.23) then makes it clear that, while displacement vector x^μ is naturally contravariant as in (10.19) and (10.21), the del operator is naturally covariant. We shall often use the notation ∂_μ to represent this covariant del operator:

$$\partial_\mu \equiv \frac{\partial}{\partial x^\mu} = \left(\frac{1}{c}\frac{\partial}{\partial t}, \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right), \tag{10.26}$$

and the corresponding contravariant del-operator

$$\partial^\mu = \eta^{\mu\nu}\partial_\nu = \left(-\frac{1}{c}\frac{\partial}{\partial t}, \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right). \tag{10.27}$$

A contraction of the two operators in (10.26) and (10.27) leads to the Lorentz-invariant 4-Laplacian (D'Alembertian) operator:

$$\Box \equiv \partial^\mu \partial_\mu = -\frac{1}{c^2}\frac{\partial^2}{\partial t^2} + \nabla^2,$$

with the Laplacian operator being $\nabla^2 = (\partial^2/\partial x^2) + (\partial^2/\partial y^2) + (\partial^2/\partial z^2)$. Thus the relativistic wave equation has the form of $\Box\psi = 0$.

The velocity 4-vector

We have already shown in Chapter 2 (cf. (2.21)–(2.24)) that velocity components have a rather complicated Lorentz transformation property. This is because ordinary velocity dx^μ/dt is not a proper 4-vector: while dx^μ is a 4-vector, the ordinary time coordinate t is not a Lorentz scalar—it is a component of a 4-vector. This suggests that to have a velocity satisfying simple Lorentz transformation, we should differentiate the displacement with respect to the proper time τ, which is a Lorentz scalar:

$$U^\mu = \frac{dx^\mu}{d\tau}. \tag{10.28}$$

Still, the relation between the 4-velocity U^μ and dx^μ/dt can be readily deduced. Coordinate time and proper time being related by the time dilation relation of

$$t = \gamma\tau \tag{10.29}$$

with

$$\gamma = \left(1 - \frac{v^2}{c^2}\right)^{-1/2} \qquad \text{with } v_i = \frac{dx_i}{dt}, \tag{10.30}$$

we have

$$U^\mu = \frac{dx^\mu}{d\tau} = \gamma\frac{dx^\mu}{dt} = \gamma(c, v_1, v_2, v_3). \tag{10.31}$$

It is easy to check the invariance of $U^\mu U_\mu \equiv U^2$:

$$U^2 = \gamma^2(-c^2 + v^2) = -c^2. \tag{10.32}$$

Remark: In Chapter 6, where motion in the Schwarzschild spacetime was discussed, we have used the Lagrangian $L = g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu$ where $\dot{x}^\mu = dx^\mu/d\tau$ can be interpreted as the 4-velocity of the particle. Thus $L = U^2 = -c^2$ for a material test particle (6.40). It should be noted that because there is no rest frame for the photon, the corresponding concept of 4-velocity as differentiation of displacement with respect to proper time does not exist. In that case, one can replace τ by the curve parameter of a photon's worldline; then the connection to a time coordinate is lost.

The relativistic momentum

For momentum, we naturally consider the product of invariant mass m with 4-velocity:

$$p^\mu \equiv mU^\mu = \gamma(mc, \mathbf{p}_{\rm NR}), \tag{10.33}$$

where we have used (10.31) with $\mathbf{p}_{\rm NR} = m\mathbf{v}$ being the nonrelativistic momentum. Namely, the 1, 2, 3 components of p^μ are the relativistic generalization of the ordinary momentum, $\mathbf{p} = \gamma m\mathbf{v}$. What then is the zeroth component of the 4-momentum? Let's take its nonrelativistic limit:

$$p^0 = mc\gamma = mc\left(1 - \frac{v^2}{c^2}\right)^{-1/2} \xrightarrow{\text{NR}} mc\left(1 + \frac{1}{2}\frac{v^2}{c^2} + \cdots\right)$$
$$= \frac{1}{c}\left(mc^2 + \frac{1}{2}mv^2 + \cdots\right). \qquad (10.34)$$

The presence of the kinetic energy term $\frac{1}{2}mv^2$ in the nonrelativistic limit naturally suggests that we interpret cp^0 as the relativistic energy $E = \gamma mc^2$, which has a nonvanishing value mc^2 even when the particle is at rest $v = 0$:

$$p^\mu = \left(\frac{E}{c}, \mathbf{p}\right). \qquad (10.35)$$

Because the invariant square of the 4-velocity is $-c^2$, (10.32), the invariant square of the 4-momentum must be $-(mc)^2$. From this we obtain the important relativistic energy–momentum relation:

$$E^2 = (mc^2)^2 + (\mathbf{p}c)^2. \qquad (10.36)$$

In summary, we have the relativistic 3-momentum $\mathbf{p} = \gamma m\mathbf{v}$, and the relativistic energy $E = \gamma mc^2$, which encompasses the rest ($\gamma = 1$) energy of

$$E = mc^2. \qquad (10.37)$$

Remark: The concept of a velocity-dependent mass $m^* \equiv \gamma m$ is sometimes used in the literature so that $\mathbf{p} = m^*\mathbf{v}$ and $E = m^*c^2$. In our discussion we will avoid such a usage and restrict ourselves only to the Lorentz scalar mass m, which is equal to m^* in the rest frame of particle $m^*|_{v=0} = m$—hence called the **rest mass**.

Covariant force

Just as the ordinary velocity $\mathbf{v}$ has a complicated Lorentz property and we introduced the object of 4-velocity, it is also not easy to relate different components of the usual force vector $\mathbf{F} = d\mathbf{p}/dt$ in different moving frames. The notions of 4-velocity and 4-momentum naturally lead us to the definition of 4-force, or the covariant force, as

$$K^\mu \equiv \frac{dp^\mu}{d\tau} = m\frac{dU^\mu}{d\tau}, \qquad (10.38)$$

which, using (10.35), has components

$$K^\mu = \frac{dp^\mu}{d\tau} = \gamma\frac{d}{dt}\left(\frac{E}{c}, \mathbf{p}\right) = \gamma\left(\frac{\dot{E}}{c}, \mathbf{F}\right). \qquad (10.39)$$

Next we show that the rate of energy change $\dot{E}$ is given, just as in nonrelativistic physics, by the dot-product $\mathbf{F} \cdot \mathbf{v}$. From (10.38)

$$K^\mu U_\mu = m\frac{dU^\mu}{d\tau}U_\mu = \frac{1}{2}m\frac{d}{d\tau}U^\mu U_\mu = 0, \qquad (10.40)$$

because U^2 is a constant. Substituting in the components of K^μ and U_μ from (10.39) and (10.31), we have

$$K^\mu U_\mu = \gamma^2(-\dot{E} + \mathbf{F} \cdot \mathbf{v}) = 0. \qquad (10.41)$$

Hence the components of the covariant force can be displayed,

$$K^{\mu} = \gamma\left(\frac{\mathbf{F}\cdot\mathbf{v}}{c}, \mathbf{F}\right). \tag{10.42}$$

In Box 10.1 we discuss another familiar 4-vector, with frequency and wave number as its components.

Box 10.1 The wave vector

Recall that for a dynamic quantity $A(\mathbf{x}, t)$ to be a solution to the wave equation, its dependence on the space and time coordinates must be in the combination of $(\mathbf{x} - \mathbf{v}t)$, where $\mathbf{v}$ is the wave velocity. A harmonic electromagnetic wave is then proportional to $\exp i(\mathbf{k}\cdot\mathbf{x} - \omega t)$ with $k = |\mathbf{k}| = 2\pi/\lambda$ being the wave number, $\omega = 2\pi/T$ being the angular frequency, and they being related to the light velocity as $\omega/k = c$.

The phase factor $(\mathbf{k}\cdot\mathbf{x} - \omega t)$, basically counting the number of peaks and troughs of the wave, must be a frame-independent quantity, that is, a Lorentz scalar,

$$\mathbf{k}\cdot\mathbf{x} - \omega t = (-ct, \quad \mathbf{x})\begin{pmatrix}\omega/c\\ \mathbf{k}\end{pmatrix} \equiv x_{\mu}k^{\mu}.$$

From our knowledge that $(ct, \mathbf{x})$ is a 4-vector and $x_{\mu}k^{\mu}$ a scalar, we conclude (via the quotient theorem, Problem 10.6) that ω and $\mathbf{k}$ must also form a 4-vector:

$$k^{\mu} = \left(\frac{\omega}{c}, k_x, k_y, k_z\right). \tag{10.43}$$

Namely, under the Lorentz transformation, the components of this wave vector change into each other as

$$k^{\mu} \longrightarrow k'^{\mu} = [\mathbf{L}]^{\mu}_{\nu}k^{\nu}. \tag{10.44}$$

Specifically under a Lorentz boost in the +x direction with reduced velocity $\beta = v/c$, we have (cf. Eq. (10.20))

$$k'_x = \gamma\left(k_x - \beta\frac{\omega}{c}\right), \tag{10.45}$$

$$\omega' = \gamma(\omega - c\beta k_x) = \gamma(\omega - c\beta k\cos\theta), \tag{10.46}$$

where θ is the angle between the boost direction $\hat{\mathbf{x}}$ and the direction of wave propagation $\hat{\mathbf{k}}$. In this way we have the **relativistic Doppler formula**,

$$\omega' = \frac{(1 - \beta\cos\theta)}{\sqrt{1 - \beta^2}}\omega, \tag{10.47}$$

which is to be compared to the nonrelativistic Doppler relation $\omega' = (1 - \beta\cos\theta)\omega$. We note that in the NR limit there is no Doppler shift in the transverse direction $\theta = \pi/2$—compared to the relativistic "transverse Doppler effect" of $\omega' = \gamma\omega$. (One can trace back the origin of this new effect as due to the SR time dilation effect.) In the longitudinal direction $\theta = 0$ we have the familiar relation of

$$\frac{\omega'}{\omega} = \sqrt{\frac{1 - \beta}{1 + \beta}}. \tag{10.48}$$

Because of the $\omega = ck$ relation k^{μ} has a nil invariant length, $k^{\mu}k_{\mu} = 0$.

10.3 Manifestly covariant formalism for E&M

It has already been shown in Chapter 2 (cf. Box 2.1) that electromagnetic theory respects the principle of relativity. In this section, we will present the equations of electromagnetism in "manifestly covariant form." Namely, they will be written in a form making it obvious that these relations do not change under Lorentz transformations.

10.3.1 The electromagnetic field tensor

Relativity unifies space and time, and space and time coordinates become components of a common vector in the covariant formalism. They can be transformed into each other when viewed in different inertial frames. Relativity also makes clear the unification of electricity and magnetism, as **E** and **B** fields can be transformed into each other by Lorentz transformations, (2.18), and they must be elements belonging to the same tensor (Box 10.2). The six fields, E_i and B_i, $(i = 1, 2, 3)$, are independent elements of a common antisymmetric $F_{\mu\nu} = -F_{\nu\mu}$ 4-field tensor, with the assignment of

$$F_{0i} = -E_i, \quad F_{ij} = \varepsilon_{ijk} B_k, \tag{10.49}$$

where ε_{ijk} is the totally antisymmetric Levi–Civita symbol with $\varepsilon_{123} = 1$. Explicitly writing out (10.49), we have

$$F_{\mu\nu} = \begin{pmatrix} 0 & -E_1 & -E_2 & -E_3 \\ E_1 & 0 & B_3 & -B_2 \\ E_2 & -B_3 & 0 & B_1 \\ E_3 & B_2 & -B_1 & 0 \end{pmatrix} \tag{10.50}$$

or

$$F^{\mu\nu} = g^{\mu\lambda} F_{\lambda\rho} g^{\rho\nu} = \begin{pmatrix} 0 & E_1 & E_2 & E_3 \\ -E_1 & 0 & B_3 & -B_2 \\ -E_2 & -B_3 & 0 & B_1 \\ -E_3 & B_2 & -B_1 & 0 \end{pmatrix}. \tag{10.51}$$

It is also useful to define the **dual field tensor**

$$\tilde{F}_{\mu\nu} \equiv -\frac{1}{2} \varepsilon_{\mu\nu\lambda\rho} F^{\lambda\rho}, \tag{10.52}$$

where $\varepsilon_{\mu\nu\lambda\rho}$ is the 4D Levi–Civita symbol[2] with $\varepsilon_{0ijk} = \varepsilon_{ijk}$ and thus $\varepsilon_{0123} = 1$. Namely,

$$\tilde{F}_{0i} = -B_i, \quad \tilde{F}_{ij} = -\varepsilon_{ijk} E_k \tag{10.53}$$

or explicitly

$$\tilde{F}_{\mu\nu} = \begin{pmatrix} 0 & -B_1 & -B_2 & -B_3 \\ B_1 & 0 & -E_3 & E_2 \\ B_2 & E_3 & 0 & -E_1 \\ B_3 & -E_2 & E_1 & 0 \end{pmatrix}. \tag{10.54}$$

Box 10.2 Lorentz transformation of the EM fields

With respect to moving observers, the electric and magnetic fields transform into each other. That they are components of a 4-tensor means that they

[2] The Levi–Civita symbol in an n-dimensional space is a quantity with n indices. Thus, in a 3D space we have ϵ_{ijk} with $(i = 1, 2, 3)$, and in a 4D space, $\epsilon_{\mu\nu\lambda\rho}$, etc. They are totally antisymmetric: an interchange of any two indices results in a minus sign: $\epsilon_{ijk} = -\epsilon_{jik} = \epsilon_{jki}$, etc. Thus, they vanish whenever any two indices are equal. All the nonzero elements can be obtained by permutation of indices from $\epsilon_{12} = \epsilon_{123} = \epsilon_{0123} \equiv 1$. Among their utilities, they can be used to express the cross-product of vectors, $(\mathbf{A} \times \mathbf{B})_i = \epsilon_{ijk} A_j B_k$. For further discussion, see Section 11.3.

must transform according to (10.11), as:

$$F^{\mu\nu} \longrightarrow F'^{\mu\nu} = [\mathbf{L}]^{\mu}_{\lambda}[\mathbf{L}]^{\nu}_{\rho}F^{\lambda\rho}. \tag{10.55}$$

The explicit form of this transformation under a boost (see (10.20)) is given by

$$\begin{pmatrix} 0 & E_1' & E_2' & E_3' \\ -E_1' & 0 & B_3' & -B_2' \\ -E_2' & -B_3' & 0 & B_1' \\ -E_3' & B_2' & -B_1' & 0 \end{pmatrix} = \begin{pmatrix} \gamma & -\beta\gamma & & \\ -\beta\gamma & \gamma & & \\ & & 1 & \\ & & & 1 \end{pmatrix}$$

$$\times \begin{pmatrix} 0 & E_1 & E_2 & E_3 \\ -E_1 & 0 & B_3 & -B_2 \\ -E_2 & -B_3 & 0 & B_1 \\ -E_3 & B_2 & -B_1 & 0 \end{pmatrix} \begin{pmatrix} \gamma & -\beta\gamma & & \\ -\beta\gamma & \gamma & & \\ & & 1 & \\ & & & 1 \end{pmatrix}. \tag{10.56}$$

One can easily check that the relation given by the tensor transformation (10.56) is just the Lorentz transformation $(\mathbf{E}, \mathbf{B}) \rightarrow (\mathbf{E}', \mathbf{B}')$ as shown in (2.18).

The dual tensor $\tilde{F}_{\mu\nu}$ has the same Lorentz transformation property as the tensor $F_{\mu\nu}$ itself.[4] From this we deduce that under Lorentz transformation there are **two** combinations of the electromagnetic fields that are invariant (products among $F^{\mu\nu}$ and $\tilde{F}_{\mu\nu}$ with all indices contracted):

$$F^{\mu\nu}F_{\mu\nu} \propto (\mathbf{E}^2 - \mathbf{B}^2) \tag{10.57}$$

and

$$F^{\mu\nu}\tilde{F}_{\mu\nu} \propto (\mathbf{E} \cdot \mathbf{B}). \tag{10.58}$$

NB: The combination $\mathbf{E}^2 + \mathbf{B}^2$ is not a Lorentz scalar. It, being the energy density, transforms as a component of the energy–momentum tensor, which we shall discuss in Box 10.5.

We now use the field tensor $F_{\mu\nu}$ to write the equations of electromagnetism in a form that clearly displays their Lorentz covariance.

Lorentz force law

Using (10.49) one can easily show that the electromagnetic equation of motion

$$\mathbf{F} = q\left(\mathbf{E} + \frac{1}{c}\mathbf{v} \times \mathbf{B}\right) \tag{10.59}$$

is just the covariant Lorentz force law

$$K^{\mu} = \frac{q}{c}F^{\mu\nu}U_{\nu}. \tag{10.60}$$

with μ taking on the spatial indices $(1, 2, 3)$. The interpretation of $\mu = 0$ is left as an exercise.

[4] $F_{\mu\nu}$ and $\tilde{F}_{\mu\nu}$ transform differently under space reflection. As a result, $(\mathbf{E}^2 - \mathbf{B}^2)$ is a scalar and $(\mathbf{E} \cdot \mathbf{B})$ a pseudoscalar. Also we have $F^{\mu\nu}F_{\mu\nu} = \tilde{F}^{\mu\nu}\tilde{F}_{\mu\nu}$.

Inhomogeneous Maxwell's equations

Gauss's and Ampere's laws in (2.17)

$$\nabla \cdot \mathbf{E} = \rho, \quad \nabla \times \mathbf{B} - \frac{1}{c}\frac{\partial \mathbf{E}}{\partial t} = \frac{\mathbf{j}}{c}, \tag{10.61}$$

are contained in the covariant Maxwell's equation

$$\partial_\mu F^{\mu\nu} = -\frac{1}{c} j^\nu, \tag{10.62}$$

where the electromagnetic 4-current is given as

$$j^\mu = (j^0, \mathbf{j}) = (c\rho, \mathbf{j}). \tag{10.63}$$

Homogeneous Maxwell's equations

Faraday's and magnetic-Gauss's laws in (2.16)

$$\nabla \times \mathbf{E} + \frac{1}{c}\frac{\partial \mathbf{B}}{\partial t} = 0, \quad \nabla \cdot \mathbf{B} = 0 \tag{10.64}$$

are contained in the "Bianchi identity" of the EM field tensor (Box 10.3)

$$\partial_\mu F_{\nu\lambda} + \partial_\lambda F_{\mu\nu} + \partial_\nu F_{\lambda\mu} = 0. \tag{10.65}$$

Alternatively, the homogeneous equations can be written as (Problem 10.12)

$$\partial_\mu \tilde{F}^{\mu\nu} = 0. \tag{10.66}$$

Remark: Written in (10.62) and (10.66), it is clear that Maxwell's equations in the free space (i.e. $j^\mu = 0$) are invariant under the **duality transformation** of $F_{\mu\nu} \to \tilde{F}_{\mu\nu}$, namely, a 90° rotation in the plane spanned by perpendicular **E** and **B** axes: $\mathbf{E} \to \mathbf{B}$ and $\mathbf{B} \to -\mathbf{E}$.

Box 10.3 EM potential and gauge symmetry

The homogeneous Maxwell's Eq. (10.66) $\epsilon^{\mu\nu\lambda\rho}\partial_\mu F_{\lambda\rho} = 0$ can be solved by expressing the field tensor in terms of the electromagnetic 4-potential A^μ:

$$F_{\lambda\rho} = \partial_\lambda A_\rho - \partial_\rho A_\lambda. \tag{10.67}$$

Equation (10.67) is just the familiar relation between EM fields and the scalar and vector potentials $(\phi, \mathbf{A}) = A^\mu$:

$$\mathbf{B} = \nabla \times \mathbf{A}, \quad \mathbf{E} = -\nabla\phi - \frac{\partial}{\partial t}\mathbf{A}. \tag{10.68}$$

The dynamics of the EM potentials is then determined by the inhomogeneous Maxwell's Eqs (10.62) after replacement of (10.67). This simplifies the description by reducing the number of dynamical variables from six in $F_{\mu\nu}$ down to four A_μ. However, the correspondence between $F_{\mu\nu}$ and A_μ is not unique because of the relation (10.67), hence the Maxwell's equations are invariant under the "**gauge transformation**"

$$A_\mu \longrightarrow A'_\mu = A_\mu - \partial_\mu \chi, \tag{10.69}$$

where $\chi(x)$ is an arbitrary spacetime dependent scalar function, (called the gauge function). A change of the potential according to (10.69) will not alter the electromagnetic description by **E** and **B** fields.

10.3.2 Electric charge conservation

When the 4-del operator ∂_ν is applied to the inhomogeneous Maxwell's Eq. (10.62), the left-hand side (LHS) vanishes because the combination $\partial_\nu \partial_\mu$ is symmetric in (μ, ν) while $F^{\mu\nu}$ is antisymmetric. This implies that the right-hand side (RHS) must also vanish:

$$\partial_\nu j^\nu = 0. \tag{10.70}$$

To investigate the physical meaning of this 4-divergence equation, we display its components as shown in (10.63):

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{j} = 0, \tag{10.71}$$

which is the familiar "equation of continuity." If we integrate every term over the volume,

$$\frac{d}{dt} \int_V \rho dV = - \int_V \nabla \cdot \mathbf{j} dV = - \oint_S \mathbf{j} \cdot d\boldsymbol{\sigma}, \tag{10.72}$$

where we have used the divergence theorem to arrive at the last integral (over the closed surface S, covering the volume V). This expression clearly shows the physical interpretation of this equation as a statement of electric charge conservation: the RHS shows the in-flow of electric charge across the surface S (flux) resulting in an increase of charge in the volume V, as expressed on the LHS. As a general rule, the expression of any conservation laws in systems of continuous media (e.g. a field system) is in the form of continuity Eq. (10.71), or more directly as the vanishing 4-divergence condition (10.70).

10.4 Energy–momentum tensors

We have just studied the electromagnetic current

$$j^\mu = (c\rho, \mathbf{j}), \tag{10.73}$$

where ρ is the electric charge and $\mathbf{j}$ the current density. More explicitly the x-component can be written out as

$$j_1 = \frac{\Delta q}{(\Delta y \Delta z)\Delta t} = \rho v_x. \tag{10.74}$$

We have the relation between charge–current–density and charge–density as $\mathbf{j} = \rho \mathbf{v}$, where $\mathbf{v}$ is the velocity field. Thus, (10.73) may be written as $j^\mu = \rho(c, \mathbf{v})$. We can also replace the density by the rest frame density ρ' (which is a Lorentz scalar) through the relation $\rho = \gamma \rho'$ (reflecting the usual Lorentz length/volume contraction) to relate it to the 4-velocity field,

$$j^\mu = \rho' \gamma(c, \mathbf{v}) = \rho' U^\mu. \tag{10.75}$$

This shows explicitly that, ρ' being a scalar, j^μ is a *bona fide* 4-vector.

Electric charge conservation can be expressed as a 4-divergence condition (10.70). Other conservation laws can all be written similarly. For example, instead of charge, if we consider the case of mass, we can similarly define a mass-current 4-vector as (10.73):

$$j^\mu = (c\rho, \mathbf{j}) = (c \times \text{mass density, mass current-density}) \tag{10.76}$$

and (10.70) becomes a statement of mass conservation.

In the same manner, we can consider a 4-current for energy ($p^0 = E/c$):

$$J^{(0)\mu} = \left(\text{energy density, } \frac{1}{c} \times \text{energy current–density}\right), \tag{10.77}$$

as well as the 4-current for the ith momentum component (p^i):

$$J^{(i)\mu} = (c \times p^i \text{ density}, p^i \text{ current–density}). \tag{10.78}$$

The factor of c in the last line originates from the current normalization: $(c\rho, \mathbf{j})$. These four 4-currents, $J^{(0)\mu}$ and $J^{(i)\mu}$ (with $i = 1, 2, 3$), are not independent because, unlike charge and mass, energy and momentum components transform into each other under a Lorentz transformation. That is, they themselves form a 4-vector: $(p^0, p^i) = p^\mu$. This suggests that we need to place these four ($\nu = 0, i$) 4-currents $J^{(\nu)\mu}$ ($\mu = 0, j$) together in one multiplet, in the form of a 4×4 matrix:

$$T^{\nu\mu} = \begin{pmatrix} J^{(0)\mu} \\ J^{(i)\mu} \end{pmatrix} = \begin{pmatrix} J^{(0)0}, J^{(0)j} \\ J^{(i)0}, J^{(i)j} \end{pmatrix}, \tag{10.79}$$

called the (symmetric) **energy–momentum tensor** $T^{\mu\nu}$. We are particularly interested in this quantity because energy and momentum being the source of gravity, $T^{\mu\nu}$ enters directly in the the relativistic field equation of gravity. (This Einstein equation, first displayed in Section 5.3.2, will be discussed in detail in Section 12.2.) Here we first examine the physical meaning of each component of this tensor $T^{\mu\nu}$:

- $T^{00} = J^{(0)0} =$ energy density, cf. (10.77).
- $T^{0i} = J^{(0)i} =$ ith-component of $((1/c)$ of energy) current–density, cf. (10.77). For example,
$$T^{01} = \frac{\gamma mc^2/c}{(\Delta y \Delta z)\Delta t} = \frac{mc}{\Delta V}\gamma\frac{\Delta x}{\Delta t} = \rho' c\gamma v_x = \frac{\rho' c^2}{c} U_x. \tag{10.80}$$
- $T^{i0} = J^{(i)0} = c \times$ density of ith-momentum component, cf. (10.78). For example,
$$T^{10} = c\frac{\gamma m v_x}{\Delta V} = \rho' c U_x. \tag{10.81}$$
We see that momentum density T^{i0} is equal to the energy current T^{0i}.
- $T^{ij} = J^{(i)j} =$ the jth component of the ith-momentum current density. We note that the diagonal $i = j$ momentum current T^{ii} has the simple interpretation of being the pressure:
$$\text{momentum–current} = \frac{\text{momentum}}{(\text{area})_\perp \Delta t} = \frac{\text{force}}{(\text{area})_\perp} = \text{pressure}, \tag{10.82}$$
where we have used the fact the rate of momentum change is force. Clearly, the off-diagonal terms would involve shear forces. Also, just as $T^{i0} = T^{0i}$, it is straightforward to show that $T^{ij} = T^{ji}$. Thus, in general,
$$T^{\nu\mu} = T^{\mu\nu}. \tag{10.83}$$
$T^{\nu\mu}$ is a symmetric tensor.

Energy momentum conservation, for an isolated system, is then expressed as

$$\partial_\mu T^{\mu\nu} = 0. \tag{10.84}$$

If the system is subject to some external force field, the RHS will then be the force density field,

$$\partial_\mu T^{\mu\nu} = \phi^\nu. \tag{10.85}$$

This is analogous to the single particle case where the change of momentum is related to force: $dp^\mu/d\tau = K^\mu$; in the absence of any external force, we have momentum conservation: $dp^\mu/d\tau = 0$.

A swarm of noninteracting particles

Let us construct the energy–momentum tensor for the simplest system of a swarm of noninteracting particles, a "cloud of dust." First consider a coordinate system, in which each position label is carried by the particles themselves. In such **comoving frames**, all particles have a fixed position-coordinate at all times, thus with respect to this coordinate system all particles are effectively at rest ($v = 0, \gamma = 1$), the 4-velocity field takes on a simple form of $U^\mu = (c, \mathbf{0})$. The only nonvanishing energy–momentum tensor term is the rest energy density T^{00}:

$$T^{\mu\nu} = \begin{pmatrix} \rho' c^2 & & & \\ & 0 & & \\ & & 0 & \\ & & & 0 \end{pmatrix}, \tag{10.86}$$

where ρ' is the mass density in this comoving (i.e. rest) coordinate system. But in this reference frame we can also write the above tensor in terms of the 4-velocity field $U^\mu = (c, \mathbf{0})$ as

$$T^{\mu\nu} = \rho' U^\mu U^\nu. \tag{10.87}$$

Even though we have arrived at this expression for $T^{\mu\nu}$ in a particular coordinate frame, because this equation is a proper tensor equation (i.e. every term has the same tensor property), it is covariant under coordinate transformations. Consequently this expression is also valid in **every** inertial frame, and hence is the general expression of the energy–momentum tensor for this system of noninteracting particles.

Ideal fluid

We now consider the case of an **ideal fluid**, in which fluid elements interact only through a normal (perpendicular) force. Namely, there is no shear force (thus $T^{ij} = 0$ for $i \neq j$). This implies that in this system there is pressure but no viscosity. So to obtain the $T^{\mu\nu}$ of an ideal fluid from that in (10.86), all we need to do is to add pressure terms (cf. (10.82)) in the (i, i) diagonal positions:

$$T^{\mu\nu} = \begin{pmatrix} \rho' c^2 & & & \\ & p & & \\ & & p & \\ & & & p \end{pmatrix} = \left(\rho' + \frac{p}{c^2}\right) U^\mu U^\nu + p\eta^{\mu\nu}. \tag{10.88}$$

The equality of $T^{11} = T^{22} = T^{33} = p$ expresses the isotropy property of the ideal fluid. Namely, the pressure applied to a given portion of the fluid is transmitted equally in all directions and is everywhere perpendicular to the surface on which it acts. Also, because the given volume element is at rest in the comoving frame, its momentum density also vanishes, $T^{0i} = T^{i0} = 0$. Similar to the above-discussed cloud of dust case, the proper tensor expression on the RHS of (10.88) means that this expression is valid for all coordinate frames. The nonrelativistic limit of an ideal fluid's energy momentum tensor, and its relation to Euler's equation for fluid mechanics, are considered in Box 10.4.

Box 10.4 Nonrelativistic limit and the Euler's equation

It is instructive to consider the nonrelativistic limit ($\gamma \to 1$) of the energy–momentum tensor of an ideal fluid. As the rest energy dominates over pressure (which results from particle momenta) $\rho c^2 \gg p$, the tensor in (10.88) takes on the form of

$$T^{\mu\nu} \overset{\text{NR}}{=} \begin{pmatrix} \rho c^2 & \rho c v_i \\ \rho c v_j & \rho v_i v_j + p\delta_{ij} \end{pmatrix}. \tag{10.89}$$

Let us now examine the conservation law $\partial_\mu T^{\mu\nu} = 0$ for this nonrelativistic system. (In the following discussion we shall often use the expression for mass current density $\mathbf{j} = \rho\mathbf{v}$.)

- $\partial_\mu T^{\mu 0} = \partial_0 T^{00} + \partial_i T^{i0} = c(\partial\rho/\partial t + \nabla \cdot \mathbf{j}) = 0$, which is just the continuity equation, expressing mass conservation.
- $\partial_\mu T^{\mu j} = \partial_0 T^{0j} + \partial_i T^{ij} = \partial_t \rho v_j + \partial_i(\rho v_i v_j + p\delta_{ij}) = 0$, which is Euler's equation of fluid mechanics:

$$\rho\left[\frac{\partial\mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v}\right] = -\nabla p. \tag{10.90}$$

To see the physical significance of the terms in this equation, let us recall that, the pressure p being the force per unit area, the total force acting on a closed surface is given by

$$-\oint_S p d\boldsymbol{\sigma} = -\int_V \nabla p dV.$$

Thus $-\nabla p dV$ is seen as the force acting on a fluid element having a volume dV,

$$-\nabla p dV = (\rho dV)\frac{d\mathbf{v}}{dt}, \tag{10.91}$$

where $d\mathbf{v}/dt$ represents the rate of change of velocity of a fluid element as it moves about in space. Namely, here $\mathbf{v}$ is the velocity field and depends on time as well as on the spatial position:

$$\frac{d\mathbf{v}}{dt} = \frac{\partial\mathbf{v}}{\partial t} + \frac{dx}{dt}\frac{\partial\mathbf{v}}{\partial x} + \frac{dy}{dt}\frac{\partial\mathbf{v}}{\partial y} + \frac{dz}{dt}\frac{\partial\mathbf{v}}{\partial z} = \frac{\partial\mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v}. \tag{10.92}$$

Namely, Euler's Eq. (10.90) is just the "$\mathbf{F} = m\mathbf{a}$ equation" for fluid mechanics,

$$\rho\frac{d\mathbf{v}}{dt} = -\nabla p. \tag{10.93}$$

The electromagnetic field is a physical system carrying energy and momentum. In Box 10.5 we discuss the $T_{\mu\nu}$ for such a system.

Box 10.5 $T_{\mu\nu}$ of the electromagnetic field

It can be shown[5] that the energy–momentum tensor for an electromagnetic field is

$$T^{\mu\nu} = \frac{1}{2}\eta_{\alpha\beta}(F^{\mu\alpha}F^{\nu\beta} + \tilde{F}^{\mu\alpha}\tilde{F}^{\nu\beta}) = \eta_{\alpha\beta}F^{\mu\alpha}F^{\nu\beta} - \frac{1}{4}\eta^{\mu\nu}F^{\alpha\beta}F_{\alpha\beta}, \tag{10.94}$$

where we have used the identity

$$\eta_{\alpha\beta}(F^{\mu\alpha}F^{\nu\beta} - \tilde{F}^{\mu\alpha}\tilde{F}^{\nu\beta}) = \frac{1}{2}\eta^{\mu\nu}F^{\alpha\beta}F_{\alpha\beta}. \tag{10.95}$$

This relation can be proven by summing over Levi–Civita symbols appearing in the definition of the dual fields $\tilde{F}^{\mu\alpha}$ and $\tilde{F}^{\nu\beta}$, or by direct multiplication of field tensor matrices of (10.50) and (10.54). From the component expression of the field tensor, one can easily check (Problem 10.15) that $T^{00} = \frac{1}{2}(\mathbf{E}^2 + \mathbf{B}^2)$ and $T^{0i} = (\mathbf{E} \times \mathbf{B})_i$, which are, respectively, the familiar EM expressions for the energy density and the energy current density (the Poynting vector).

For the simpler case of free space $j^\mu = 0$, we expect to have conservation of field energy–momentum $\partial_\mu T^{\mu\nu} = 0$. This can be checked as follows:

$$\begin{aligned}
\partial_\mu T^{\mu\nu} &= \eta_{\alpha\beta}F^{\mu\alpha}(\partial_\mu F^{\nu\beta}) - \frac{1}{2}\eta^{\mu\nu}F_{\alpha\beta}(\partial_\mu F^{\alpha\beta}) \\
&= F_{\alpha\beta}(\partial^\alpha F^{\nu\beta}) - \frac{1}{2}F_{\alpha\beta}(\partial^\nu F^{\alpha\beta}) \\
&= F_{\alpha\beta}(\partial^\alpha F^{\nu\beta}) + \frac{1}{2}F_{\alpha\beta}(\partial^\beta F^{\nu\alpha} + \partial^\alpha F^{\beta\nu}) \\
&= \frac{1}{2}F_{\alpha\beta}(\partial^\alpha F^{\nu\beta} + \partial^\beta F^{\nu\alpha}) = 0,
\end{aligned} \tag{10.96}$$

where on the first line we have used Maxwell's Eq. (10.62) for $j^\mu = 0$, to reach the second line we have relabeled some dummy indices, to reach the third line we have used Maxwell's Eq. (10.65), to reach the fourth line we have used the antisymmetric property of $F^{\nu\beta} = -F^{\beta\nu}$, and the last equality follows from the fact that the antisymmetric $F_{\alpha\beta}$ is contracted with the combination in the parenthesis, which is symmetric in (α, β).

For the case where there are charged particles in the space so that $j^\mu \neq 0$, energy and momentum are stored in the field as well as in the motion of the charged particles,

$$T^{\mu\nu} = T^{\mu\nu}_{\text{field}} + T^{\mu\nu}_{\text{charge}}, \tag{10.97}$$

where $T^{\mu\nu}_{\text{charge}} = \rho'_{\text{mass}}U^\mu U^\nu$, with ρ'_{mass} being the proper mass density of the charged particles. It can be shown (Problem 10.16) that neither $T^{\mu\nu}_{\text{field}}$ nor $T^{\mu\nu}_{\text{charge}}$ are conserved, but their divergences mutually cancel so that $\partial_\mu T^{\mu\nu} = 0$. Thus for the system as a whole, energy and momentum are conserved.

[5] This is usually obtained by relating the energy–momentum tensor to the variations of the Lagrangian density of the field system.

Review questions

1. What are the covariant and contravariant components of a vector? Why do we need these two kinds of vector (tensor) components?
2. Write out the Lorentz transformation of coordinates $(t, \mathbf{x})$ for a boost $\mathbf{v} = v\hat{\mathbf{x}}$, and of the differential operators (∂_t, ∇).
3. Why do we say that the position 4-vector x^{μ} is naturally contravariant and the del operator ∂_{μ} is naturally covariant?
4. Contravariant and covariant vectors transform differently. How are their transformations related?
5. Write the coordinate transformation for a mixed tensor $T^{\mu}_{\nu} \longrightarrow T^{'\mu}_{\nu}$.
6. Why is dx^{μ}/dt not a 4-vector? How is the velocity 4-vector related to this dx^{μ}/dt?
7. What are the relativistic expressions of energy and 3-momentum? Derive their invariant relation.
8. What does one mean by saying that the inhomogeneous Maxwell's equation $\partial_{\mu}F^{\mu\nu} = -j^{\nu}/c$ is manifestly covariant? Show that this equation also includes the statement of electric charge conservation.
9. What is the physical interpretation of the components in $T_{\mu\nu}$?
10. Write out the elements of $T^{\mu\nu}$ for an ideal fluid in the rest frame of a fluid element (the comoving frame).

Problems

(10.1) **Basis and inverse basis vectors: a simple exercise** Basis vectors for a two-dimensional space is given explicitly as

$$\mathbf{e}_1 = a\begin{pmatrix}1\\0\end{pmatrix} \quad \text{and} \quad \mathbf{e}_2 = b\begin{pmatrix}\cos\theta\\ \sin\theta\end{pmatrix}$$

(a) Find the inverse basis vectors $\{\mathbf{e}^i\}$ so that $\mathbf{e}_i \cdot \mathbf{e}^j = \delta_{ij}$

(b) Write out the metric tensors g_{ij} and g^{ij} so that $\sum_j g_{ij}g^{jk} = \delta_{ik}$

(c) Show that the sum of outer products is the identity matrix, $\sum_i \mathbf{e}_i \otimes \mathbf{e}^i = \mathbf{1}$. This is an expression of the completeness condition of (10.2).

(10.2) **Perpendicular vs. parallel projections** In a coordinate system with nondiagonal unit base vectors: $e_1^2 = e_2^2 = 1$, and $e_1 \cdot e_2 = \cos\alpha$, as shown in Fig. 10.1, you can use the matrix form (2.34) for the metric to check (geometrically) the relations (10.8) between the perpendicular and parallel projections as drawn in Fig. 10.1.

$$V_1 = g_{11}V^1 + g_{12}V^2 = V^1 + (\cos\alpha)V^2. \qquad (10.98)$$

(10.3) **Coordinate transformations and permutation symmetry** If a tensor has some symmetry properties, for example, $T_{ij} \pm T_{ji} = 0$, after a coordinate transformation, the transformed tensor still has the same properties; in this case, $T'_{ij} \pm T'_{ji} = 0$.

(10.4) **Transformations: components vs. basis vectors** The reason that V^i are called the contravariant components and V_i the covariant components is that they transform "oppositely" and "in the same way" as the base vectors $\mathbf{e}_i$. From the definitions given in (10.5), explain why there must be such relations.

(10.5) **g_{ij} is a tensor** We have called $g_{ij} = \mathbf{e}_i \cdot \mathbf{e}_j$ a tensor. Demonstrate that (a) the metric definition does imply the requisite transformation property; (b) use the role played by metrics in the contractions $U^iV^jg_{ij}$ or $U_iV_jg^{ij}$ to confirm these transformation properties.

(10.6) **The quotient theorem** This theorem states that in a tensor equation such as $A_{\mu\nu} = C_{\mu\lambda\rho}B^{\lambda\rho}_{\nu}$, if we know that $A_{\mu\nu}$ and $B^{\lambda\rho}_{\nu}$ are tensors, then the coefficients $C_{\mu\lambda\rho}$ must also form a tensor. Show that the proof, that g^{ij} is a set of tensor components, in Problem 10.5(b) is an illustration of the quotient theorem.

(10.7) **Lorentz transform and velocity addition rule** Use the Lorentz transformation property of the velocity 4-vector to derive the velocity addition rule (2.24).

(10.8) **Gravitational redshift: another derivation** Instead of considering a spaceship in free fall, one can use the equivalence of the spaceship at rest in a gravitational field $-\mathbf{g}$ to a spaceship moving upward with an acceleration $\mathbf{a} = \mathbf{g}$. Use the Lorentz frequency transformation of SR as given in (10.48) to derive the gravitational frequency shift (3.24) by noting that the receiver will be an observer in motion.

(10.9) **Antiproton production threshold** Because of baryon number conservation, the simplest reaction to produce an antiproton $\bar{p}$ in a proton–proton scattering is $pp \rightarrow ppp\bar{p}$. Knowing that the rest energy of a proton $m_{\text{p}}c^2 = 9.4$ GeV, use the invariant $p^\mu p_\mu$ to find the minimum kinetic energy a proton must have in order to produce an antiproton after colliding with another proton at rest.

(10.10) **Covariant Lorentz force law** Check that the $\mu = 0$ component of (10.60) does have the correct interpretation as the time component of a covariant force $K^0 = \gamma \mathbf{F} \cdot \mathbf{v}/c$ as required by (10.42).

(10.11) **Manifestly covariant Maxwell's equations** Use (10.49) and (10.53) to check that components of (10.62) and (10.66) are just the Maxwell equations of (10.61) and (10.64).

(10.12) **Homogeneous Maxwell's equations** Explicitly demonstrate that the two forms of the homogeneous Eqs (10.65) and (10.66) are equivalent. **Suggestion**: Start with Eq. (10.66) with $\nu = 0$ to derive Eq. (10.65).

(10.13) **Electromagnetic potentials** Verify the solution (10.67) of the homogeneous Maxwell equation, by substituting it into (10.65).

(10.14) $T^{\mu\nu}$ **for a swarm of dust** Use the explicit form of $\rho' U^\mu U^\nu$ in (10.87) for $T^{\mu\nu}$ to check the physical meaning of the elements of the energy–momentum tensor as discussed in the text.

(10.15) $T^{\mu\nu}$ **for electromagnetic field** Check the physical meaning of the elements of the energy–momentum tensor (10.94) for electromagnetic fields as given in the discussion prior to (10.83).

(10.16) $T^{\mu\nu}$ **for a system of EM field and charges** Show that neither $T^{\mu\nu}_{\text{field}}$ of (10.94) nor $T^{\mu\nu}_{\text{charge}}$ of (10.97) is conserved, but their divergences mutually cancel so that $\partial_\mu T^{\mu\nu} = 0$. Thus, for the system as a whole, energy and momentum are conserved.

(10.17) **Radiation pressure and energy density** Derive the relation, $p = \rho c^2/3$, between pressure and energy density for a volume of radiation by treating the radiation as an ideal fluid of (10.88). **Hint**: First examine the trace $(\eta_{\mu\nu} T^{\mu\nu})$ of the energy momentum tensor for the EM field (10.94).

Tensors in general relativity

11

- Coordinate transformations in a curved space are necessarily position-dependent. Still, the tensors used in general relativity (GR) are basically the same as those in special relativity (SR), except when differentiation is involved.
- By adding another term (related to Christoffel symbols) to the ordinary derivative operator, we can form a "covariant derivative," which produces proper tensor derivatives.
- The relation between Christoffel symbols and first derivatives of metric functions Eq. (5.10) is re-established.
- Using the concept of parallel transport, the geometric meaning of covariant differentiation is further clarified.
- The curvature tensor for an *n*-dimensional space is derived by the parallel transport of a vector around a closed path.
- Symmetry and contraction properties of the Riemann curvature tensor are considered. We find just the desired tensor needed for GR field equation.

In contrast to the case for flat space, basis vectors in a curved space must change from position to position. This implies that coordinate transformations must necessarily be position-dependent. As a consequence, ordinary derivatives of tensors, except for the trivial scalars, are no longer tensors. Nevertheless it can be shown that one can construct "covariant differentiation operations" so that they result in tensor derivatives. We demonstrate this first by formal manipulation (Section 11.1) and also by a more geometric introduction (Section 11.2). This geometric concept of parallel transport will also be employed to generalize the Gaussian curvature of a two-dimensional (2D) space to the Riemann curvature tensor for a curved space of arbitrary dimensions. We conclude this section with a study of the symmetry and contraction properties of the Riemann tensor, which will be needed when we study the GR field equation, the Einstein equation, in Chapter 12.

11.1 Derivatives in a curved space

The tensors used in general relativity (GR) are basically the same as those in special relativity (SR), except when differentiation is involved. This difference reflects the fact that coordinate transformations in a curved space are necessarily position-dependent. One finds that differentiation of a tensor results in a quantity which is no longer a tensor. This poses serious problem as relativistic equations must be tensor equations. To overcome this, we introduce in this section the "covariant derivative" which does not spoil the tensor properties, and allows us to have relativistic physics equations.

11.1.1 General coordinate transformations

The coordinate transformations in SR (the Lorentz transformations) are position-independent "global transformations." The rotation angles and boost velocity are the same for every spacetime point. We rotate the same amount of angle and boost with the same velocity everywhere. In GR we must deal with position-dependent "local transformations," the general coordinate transformation. This position dependence is related to the fact that in a curved space the basis vectors $\{\mathbf{e}_\mu\}$ must necessarily change from point to point, leading to position-dependent metric functions:

$$g_{\mu\nu} \equiv [\mathbf{e}_\mu(x) \cdot \mathbf{e}_\nu(x)] = g_{\mu\nu}(x). \tag{11.1}$$

The metric $[\mathbf{g}]$ is a rank-2 tensor and thus transforms (cf. (10.12)) as $[\mathbf{g}'] = [\mathbf{L}][\mathbf{L}][\mathbf{g}]$ where we have symbolically represented the (inverse) coordinate transformation by $[\mathbf{L}]$. If we differentiate both sides of this relation, we get

$$\partial[\mathbf{g}'] = 2[\mathbf{L}][\mathbf{g}](\partial[\mathbf{L}]) + [\mathbf{L}][\mathbf{L}](\partial[\mathbf{g}]). \tag{11.2}$$

For a flat space, one can always work with a coordinate system having a position-independent metric, $\partial[\mathbf{g}'] = \partial[\mathbf{g}] = 0$, the above relation then shows that the transformation matrix must also be position-independent, $\partial[\mathbf{L}] = 0$. In a curved space the metric must be position-dependent $\partial[\mathbf{g}] \neq 0$, implying that the transformation also has x-dependence

$$\partial[\mathbf{L}] \neq 0. \tag{11.3}$$

Coordinate transformation as a matrix of partial derivatives

The coordinate transformations in SR (the Lorentz transformations) leave invariant the separation $s^2 = g_{\mu\nu}x^\mu x^\nu$. In a curved space the bases and metric necessarily vary from point to point. General transformations in such a space are not expected to have such a finite invariant separation. However, since a curved space is locally flat, this will be possible to demand the coordinate transformation

$$dx'^\mu = [\mathbf{L}]^\mu_\nu dx^\nu \tag{11.4}$$

that leaves invariant an infinitesimal length:

$$ds^2 = g_{\mu\nu}dx^\mu dx^\nu. \tag{11.5}$$

This defines the metric functions for a given coordinate system. Let us now recall the (chain-rule) differentiation relation:

$$dx'^\mu = \frac{\partial x'^\mu}{\partial x^\nu}dx^\nu. \tag{11.6}$$

A comparison of (11.4) and (11.6) suggests that the coordinate transformation can be written as a matrix of partial derivatives:

$$[\mathbf{L}]^\mu_\nu = \frac{\partial x'^\mu}{\partial x^\nu}. \tag{11.7}$$

Namely, the transformation for a contravariant vector may be written as

$$V^{\mu} \rightarrow V'^{\mu} = \frac{\partial x'^{\mu}}{\partial x^{\nu}} V^{\nu}. \tag{11.8}$$

More explicitly, the relation in (11.8) may be written out as

$$\begin{pmatrix} V'^0 \\ V'^1 \\ V'^2 \\ V'^3 \end{pmatrix} = \begin{pmatrix} \partial x'^0/\partial x^0 & \partial x'^0/\partial x^1 & \partial x'^0/\partial x^2 & \partial x'^0/\partial x^3 \\ \partial x'^1/\partial x^0 & \partial x'^1/\partial x^1 & \partial x'^1/\partial x^2 & \partial x'^1/\partial x^3 \\ \partial x'^2/\partial x^0 & \partial x'^2/\partial x^1 & \partial x'^2/\partial x^2 & \partial x'^2/\partial x^3 \\ \partial x'^3/\partial x^0 & \partial x'^3/\partial x^1 & \partial x'^3/\partial x^2 & \partial x'^3/\partial x^3 \end{pmatrix} \begin{pmatrix} V^0 \\ V^1 \\ V^2 \\ V^3 \end{pmatrix}.$$

This notation is also applicable to the global transformation discussed in the previous chapter. As an instructive exercise, one can show the elements of the Lorentz transformation matrix (10.20) can be recovered from partial differentiation of the Lorentz boost formulae as displayed in, for example, Eq. (2.13). This way of writing the transformations also has the advantage of preventing us from misidentifying the transformation $[\mathbf{L}]^{\mu}_{\nu}$ as a tensor.

As we have discussed in Chapter 10, the del operator transforms as a covariant vector, cf. Eq. (10.25),

$$\frac{\partial}{\partial x'^{\mu}} = [\mathbf{L}]^{\nu}_{\mu} \frac{\partial}{\partial x^{\nu}} \tag{11.9}$$

The chain rule of differentiation leads to the identification

$$[\bar{\mathbf{L}}]^{\nu}_{\mu} = \frac{\partial x^{\nu}}{\partial x'^{\mu}}, \tag{11.10}$$

which makes it obvious that $[\mathbf{L}][\bar{\mathbf{L}}] = \mathbf{1}$ because $(\partial x'^{\mu}/\partial x^{\nu})(\partial x^{\lambda}/\partial x'^{\mu}) = \delta^{\lambda}_{\nu}$. For any covariant components, we have the transformation

$$V_{\mu} \rightarrow V'_{\mu} = \frac{\partial x^{\nu}}{\partial x'^{\mu}} V_{\nu}. \tag{11.11}$$

The reason $\{V_{\mu}\}$ are called **co**variant components is because they transform in the same way as the basis vectors:

$$\mathbf{e}_{\mu} \rightarrow \mathbf{e}'_{\mu} = \frac{\partial x^{\nu}}{\partial x'^{\mu}} \mathbf{e}_{\nu}, \tag{11.12}$$

while the **contra**variant components transform oppositely. A general tensor with both contravariant and covariant indices transforms as direct products of contravariant and covariant vectors $T^{\mu\nu\ldots}_{\lambda\ldots} \sim A^{\mu}B^{\nu}\ldots C_{\lambda}\ldots$. The simplest mixed tensor has the transformation

$$T^{\mu}_{\ \nu} \rightarrow T'^{\mu}_{\ \nu} = \frac{\partial x^{\lambda}}{\partial x'^{\nu}} \frac{\partial x'^{\mu}}{\partial x^{\rho}} T^{\rho}_{\ \lambda}. \tag{11.13}$$

Again we emphasize that GR coordinate transformations, that keep invariant the infinitesimal interval of Eq. (11.5), may be fruitfully viewed as the *local Lorentz transformations*—Lorentzian because they are spacetime length-preserving transformations; local because the physics equations are required to be covariant under independent transformations at every spacetime point. Also, coordinates in GR have no intrinsic significance (See discussion in Section 4.1 and 4.2 as well as in Box 6.1). Their relation to distance measurements is given through the metric function of (11.5). Thus, we are able to make coordinate changes as long as the metric is changed correspondingly. Tensors in

GR are objects that have definite transformation properties (as shown above), so that tensor equations keep their form under the general transformations. In this way these equations are automatically relativistic.

11.1.2 Covariant differentiation

The above discussion would seem to imply that there is no fundamental difference between the tensors in flat and in curved space. But as we shall demonstrate below, this is not so when differentiation is involved.

Ordinary derivatives of vector components are not tensors

In a curved space, the derivative $\partial_\nu V^\mu$ is a non-tensor. Even though we have ∂_ν and V^μ being good vectors, as indicated by (11.9),

$$\partial_\mu \to \partial'_\mu = \frac{\partial x^\lambda}{\partial x'^\mu}\partial_\lambda, \tag{11.14}$$

and (11.8), their combination $\partial_\nu V^\mu$ still does not transform properly,

$$\partial_\nu V^\mu \to \partial'_\nu V'^\mu \neq \frac{\partial x^\lambda}{\partial x'^\nu}\frac{\partial x'^\mu}{\partial x^\rho}\partial_\lambda V^\rho, \tag{11.15}$$

as required by (11.13). We can see this by differentiating $\partial'_\nu \equiv (\partial/\partial x'^\nu)$ on both sides of (11.8):

$$\partial'_\nu V'^\mu = \frac{\partial}{\partial x'^\nu}\left(\frac{\partial x'^\mu}{\partial x^\rho}V^\rho\right) = \frac{\partial x^\lambda}{\partial x'^\nu}\frac{\partial x'^\mu}{\partial x^\rho}(\partial_\lambda V^\rho) + \frac{\partial^2 x'^\mu}{\partial x'^\nu \partial x^\rho}V^\rho, \tag{11.16}$$

where (11.14) has been used. Compared to the right-hand side (RHS) of (11.15), there is an extra term

$$\frac{\partial}{\partial x'^\nu}\left(\frac{\partial x'^\mu}{\partial x^\rho}\right) \neq 0, \tag{11.17}$$

which is (11.3) with the transformation written in terms of partial derivatives. Thus, the transformation difficulty of $\partial_\nu V^\mu$ is related to the position-dependent nature of the coordinate transformation, which in turn reflects, as discussed at the beginning of this subsection, the position-dependence of the metric. Thus, the root problem lies in the moving bases $\mathbf{e}^\mu = \mathbf{e}^\mu(x)$ of the curved space. More explicitly, because the tensor components are the projections of the tensor onto the basis vectors $V^\mu = \mathbf{e}^\mu \cdot \mathbf{V}$, the moving bases $\partial_\nu \mathbf{e}^\mu \neq 0$ produce an extra term in the derivative:

$$\partial_\nu V^\mu = \mathbf{e}^\mu \cdot (\partial_\nu \mathbf{V}) + \mathbf{V} \cdot (\partial_\nu \mathbf{e}^\mu). \tag{11.18}$$

The properties of the two terms on the RHS will be studied below.

Covariant derivatives

In order for the equation to be relativistic we must have tensor equations such that they are unchanged under coordinate transformations. Thus, we seek a **covariant derivative** D_ν to be used in covariant physics equations. Such a differentiation is constructed to yield a tensor:

$$D_\nu V^\mu \to D'_\nu V'^\mu = \frac{\partial x^\lambda}{\partial x'^\nu}\frac{\partial x'^\mu}{\partial x^\rho}D_\lambda V^\rho. \tag{11.19}$$

As will be demonstrated below, the first term on the RHS of (11.18) is just this desired covariant derivative term.

We have suggested that the trouble with the differentiation of vector components is due to the coordinate dependence of V^μ. By this reasoning, derivatives of a scalar function Φ should not have this complication—because scalar tensor does not depend on the bases,

$$\partial_\mu \Phi \to \partial'_\mu \Phi' = \frac{\partial x^\lambda}{\partial x'^\mu} \partial_\lambda \Phi. \tag{11.20}$$

Similarly, the derivatives of the vector $\mathbf{V}$ itself (not its components) transform properly because $\mathbf{V}$ is coordinate-independent,

$$\partial_\mu \mathbf{V} \to \partial'_\mu \mathbf{V} = \frac{\partial x^\lambda}{\partial x'^\mu} \partial_\lambda \mathbf{V}. \tag{11.21}$$

Both (11.20) and (11.21) merely reflect the transformation of the del-operator (11.14). If we dot both sides of (11.21) by the inverse basis vectors, and use their transformation

$$\mathbf{e}'^\nu = \frac{\partial x'^\nu}{\partial x^\rho} \mathbf{e}^\rho \tag{11.22}$$

as well, we obtain

$$\mathbf{e}'^\nu \cdot \partial'_\mu \mathbf{V} = \frac{\partial x^\lambda}{\partial x'^\mu} \frac{\partial x'^\nu}{\partial x^\rho} \mathbf{e}^\rho \cdot \partial_\lambda \mathbf{V}. \tag{11.23}$$

This shows that $\mathbf{e}^\nu \cdot \partial_\mu \mathbf{V}$ is a proper mixed tensor as required by (11.13), and can be the covariant derivative we have been looking for:

$$D_\mu V^\nu = \mathbf{e}^\nu \cdot \partial_\mu \mathbf{V}. \tag{11.24}$$

This relation implies that $D_\mu V^\nu$ can be viewed as the projection of the vectors[1] $[\partial_\mu \mathbf{V}]$ along the direction of $\mathbf{e}^\nu$; we can then interpret $D_\mu V^\nu$, much in the manner of (10.5), as the coefficient of expansion of $[\partial_\mu \mathbf{V}]$ in terms of the basis vectors:

$$\partial_\mu \mathbf{V} = (D_\mu V^\nu)\mathbf{e}_\nu \tag{11.25}$$

with the repeated indices ν summed over.

Christoffel symbols as expansion coefficients of $\partial_\nu \mathbf{e}^\mu$

On the other hand, we do not have a similarly simple transformation relation like (11.21) when $\mathbf{V}$ is replaced by one of the coordinate basis vectors ($\mathbf{e}_\mu$), which by definition change under coordinate transformations. As a result, the corresponding expansion for $\partial_\nu \mathbf{e}^\mu$, as in (11.25),

$$\partial_\nu \mathbf{e}^\mu = -\Gamma^\mu_{\nu\lambda} \mathbf{e}^\lambda \quad \text{or} \quad \mathbf{V} \cdot (\partial_\nu \mathbf{e}^\mu) = -\Gamma^\mu_{\nu\lambda} V^\lambda \tag{11.26}$$

does not have coefficients $-\Gamma^\mu_{\nu\lambda}$ that are tensors. Anticipating the result, we have here used the same notation for these expansion coefficients as the **Christoffel symbols** introduced in Chapter 5 (cf. (5.10))—also called the **affine connection** (**connection**, for short).

Plugging (11.24) and (11.26) into (11.18), we find

$$D_\nu V^\mu = \partial_\nu V^\mu + \Gamma^\mu_{\nu\lambda} V^\lambda. \tag{11.27}$$

Thus, in order to produce the covariant derivative, the ordinary derivative $\partial_\nu V^\mu$ must be supplemented by another term. This second term directly reflects the position-dependence of the basis vectors, as in (11.26). Even though both $\partial_\nu V^\mu$ and $\Gamma^\mu_{\nu\lambda} V^\lambda$ do not have the correct transformation properties, the unwanted

[1] We are treating $[\partial_\mu \mathbf{V}]$ as a set of vectors, each being labeled by an index μ.

terms produced from their respective transformations (11.16) cancel each other so that their sum $D_\nu V^\mu$ is a good tensor. (Another proof of $D_\nu V^\mu$ being a tensor can be found in Problem 11.5.) Further insight about the structure of the covariant derivative can be gleaned by invoking the basic geometric concept of parallel displacement of a vector, to be presented in Section 11.2.

Compared to the contravariant vector V^μ of (11.27), the covariant derivative for a covariant vector V_μ takes on the form (Problem 11.1) of

$$D_\nu V_\mu = \partial_\nu V_\mu - \Gamma^\lambda_{\nu\mu} V_\lambda. \tag{11.28}$$

A mixed tensor such as T^μ_ν, transforming in the same way as the direct product $V^\mu U_\nu$, will have a covariant derivative

$$D_\nu T^\rho_\mu = \partial_\nu T^\rho_\mu - \Gamma^\lambda_{\nu\mu} T^\rho_\lambda + \Gamma^\rho_{\nu\sigma} T^\sigma_{\ \mu}. \tag{11.29}$$

Namely, a set of Christoffel symbols for each index of the tensor—a $(+\Gamma T)$ term for a contravariant index, a $(-\Gamma T)$ term for a covariant index, etc. A specific example is the covariant differentiation of the (covariant) metric tensor $g_{\mu\nu}$:

$$D_\lambda g_{\mu\nu} = \partial_\lambda g_{\mu\nu} - \Gamma^\rho_{\lambda\mu} g_{\rho\nu} - \Gamma^\rho_{\lambda\nu} g_{\mu\rho}. \tag{11.30}$$

11.1.3 Christoffel symbols and metric tensor

We have introduced the Christoffel symbols $\Gamma^\mu_{\nu\lambda}$ as the coefficients of expansion for $\partial_\nu \mathbf{e}^\mu$ as in (11.26). In this section, we shall relate such $\Gamma^\mu_{\nu\lambda}$ to the first derivative of the metric tensor. This will justify the identification with the symbols first defined in Eq. (5.10). To derive this relation, we need to first point out an important feature of $\Gamma^\mu_{\nu\lambda}$, as defined by (11.26) and (11.30). It can be shown (Problem 11.3)

$$\Gamma^\mu_{\nu\lambda} = \Gamma^\mu_{\lambda\nu} \tag{11.31}$$

that is, symmetric with respect to the interchange of its two lower indices.

The metric tensor is covariantly constant

While the metric tensor is position-dependent, $\partial[\mathbf{g}] \neq 0$, it is a constant with respect to covariant differentiation, $D[\mathbf{g}] = 0$ (we say $g_{\mu\nu}$ is **covariantly constant**):

$$D_\lambda g_{\mu\nu} = 0. \tag{11.32}$$

One way to prove this is to use the expression of the metric in terms of the basis vectors: $g_{\mu\nu} = \mathbf{e}_\mu \cdot \mathbf{e}_\nu$, and apply the definition of Γ, as given in (11.26), $\partial_\nu \mathbf{e}_\mu = +\Gamma^\rho_{\mu\nu} \mathbf{e}_\rho$:

$$\begin{aligned}\partial_\lambda(\mathbf{e}_\mu \cdot \mathbf{e}_\nu) &= (\partial_\lambda \mathbf{e}_\mu) \cdot \mathbf{e}_\nu + \mathbf{e}_\mu \cdot (\partial_\lambda \mathbf{e}_\nu)\\ &= \Gamma^\rho_{\lambda\mu} \mathbf{e}_\rho \cdot \mathbf{e}_\nu + \Gamma^\rho_{\lambda\nu} \mathbf{e}_\mu \cdot \mathbf{e}_\rho. \end{aligned}\tag{11.33}$$

Written in terms of the metric tensors, this relation becomes

$$\partial_\lambda g_{\mu\nu} - \Gamma^\rho_{\lambda\mu} g_{\rho\nu} - \Gamma^\rho_{\lambda\nu} g_{\mu\rho} = D_\lambda g_{\mu\nu} = 0, \tag{11.34}$$

where we have applied the definition of the covariant derivative of a covariant tensor $g_{\mu\nu}$ as in (11.30). That the metric tensor is covariantly constant is also the key ingredient in the proof of the “flatness theorem” first discussed in Section 4.2.3, and proven in Box 11.1. And as we shall see (cf. Section 12.4.3), this key property allowed Einstein to introduce his “cosmological constant term” in the general relativistic field equation.

Christoffel symbols as the metric tensor derivative

In the above discussion we have used the definition (11.26) of Christoffel symbols as the coefficients of expansion of the derivative $\partial_\nu \mathbf{e}^\mu$. Here we shall derive an expression for Christoffel symbols, as the first derivative of the metric tensor, which agrees with the definition first introduced in (5.10). We start by using several versions of (11.34) with their indices permuted cyclically:

$$\begin{aligned} D_\lambda g_{\mu\nu} &= \partial_\lambda g_{\mu\nu} - \Gamma^\rho_{\lambda\mu} g_{\rho\nu} - \Gamma^\rho_{\lambda\nu} g_{\mu\rho} = 0, \\ D_\nu g_{\lambda\mu} &= \partial_\nu g_{\lambda\mu} - \Gamma^\rho_{\nu\lambda} g_{\rho\mu} - \Gamma^\rho_{\nu\mu} g_{\lambda\rho} = 0, \\ -D_\mu g_{\nu\lambda} &= -\partial_\mu g_{\nu\lambda} + \Gamma^\rho_{\mu\nu} g_{\rho\lambda} + \Gamma^\rho_{\mu\lambda} g_{\nu\rho} = 0. \end{aligned} \tag{11.35}$$

Summing over these three equations and using the symmetry property of (11.31), we obtain:

$$\partial_\lambda g_{\mu\nu} + \partial_\nu g_{\lambda\mu} - \partial_\mu g_{\nu\lambda} - 2\Gamma^\rho_{\lambda\nu} g_{\mu\rho} = 0 \tag{11.36}$$

or, in its equivalent form,

$$\Gamma^\lambda_{\mu\nu} = \frac{1}{2} g^{\lambda\rho} [\partial_\nu g_{\mu\rho} + \partial_\mu g_{\nu\rho} - \partial_\rho g_{\mu\nu}]. \tag{11.37}$$

This relation showing $\Gamma^\mu_{\nu\lambda}$ as the first derivative of the metric tensor is called "the fundamental theorem of Riemannian geometry." It is just the definition stated previously in (5.10). From now on we shall often use this intrinsic geometric description of the Christoffel symbols (11.37) rather than (11.26). The symmetry property of (11.31) is explicitly displayed in (11.37).

Box 11.1 A proof of the flatness theorem

The flatness theorem, as first stated in Section 4.2.3, asserts that at any point P one can always make a coordinate transformation $x^\mu \to \bar{x}^\mu$ and $g^{\mu\nu} \to \bar{g}^{\mu\nu}$ where the metric tensor $\bar{g}^{\mu\nu}$ is a constant, up to a second order correction (i.e. the first order terms vanish):

$$\bar{g}^{\mu\nu}(\bar{x}) = \bar{g}^{\mu\nu}(0) + b^{\mu\nu\lambda\rho} \bar{x}_\lambda \bar{x}_\rho + \cdots , \tag{11.38}$$

where for simplicity we have taken the point P to be at the origin of the coordinate system and the position vector $\bar{x}^\mu$ is assumed to be infinitesimally small. We shall prove this result by explicit construction. Namely, we display a coordinate transformation

$$\frac{\partial x^\mu}{\partial \bar{x}^\nu} = \delta^\mu_\nu - \Gamma^\mu_{\nu\lambda} \bar{x}^\lambda \tag{11.39}$$

that is shown to lead to the result of (11.38).

Here is the proof: According to (11.39) and (11.8), the relation between the new and old coordinates can be written as $x^\mu = \bar{x}^\mu - \frac{1}{2}\Gamma^\mu_{\nu\lambda}\bar{x}^\nu\bar{x}^\lambda + \cdots$. Now, substitute (11.39), as well as the power series expansion $g^{\mu\nu}(x) = g^{\mu\nu}(0) + \partial_\lambda g^{\mu\nu} x^\lambda + \cdots$, into the metric transformation equation

$$\bar{g}_{\mu\nu}(\bar{x}) = \frac{\partial x^\lambda}{\partial \bar{x}^\mu} \frac{\partial x^\rho}{\partial \bar{x}^\nu} g_{\lambda\rho}(x), \tag{11.40}$$

we have

$$\bar{g}_{\mu\nu}(\bar{x}) = (\delta^{\lambda}_{\mu} - \Gamma^{\lambda}_{\mu\alpha}\bar{x}^{\alpha})(\delta^{\rho}_{\nu} - \Gamma^{\rho}_{\nu\beta}\bar{x}^{\beta})(g_{\lambda\rho}(0) + \partial_{\gamma} g_{\lambda\rho} x^{\gamma} + \cdots)$$
$$= g_{\mu\nu}(0) - [\Gamma^{\lambda}_{\mu\alpha} g_{\lambda\nu}(0) + \Gamma^{\lambda}_{\alpha\nu} g_{\mu\lambda}(0) - \partial_{\alpha} g_{\mu\nu}]x^{\alpha} + \cdots .$$

The coefficient of x^{α} (square bracket) vanishes because of (11.34): the metric is covariantly constant. Thus the transformation in (11.39) indeed has the claimed property of leading to a metric having the form of (11.38). ■

This proves the flatness theorem, and the assertion that $\{\bar{x}^{\mu}\}$, the local Euclidean frame (LEF), always exists. While the constant tensor can be diagonalized to the principal axes (with length adjusted correctly) so that $\bar{g}^{\mu\nu}(0)$ becomes the standard flat space metric $\eta^{\mu\nu}$, it is apparent that the second derivatives of $\bar{g}_{\mu\nu}$, related to the intrinsic curvature of the space, cannot be eliminated by adjusting the coordinate system.

Now we have shown that the transformation in (11.39) can perform the task of changing any coordinates to one which is explicitly flat in the infinitesimal region around a given point. How did one find this transformation in the first place? One can motivate the result (11.39) by comparing it to (11.27) for the case of $V^{\mu} = x^{\mu}$: the covariant derivative term, being valid in every coordinate system (including the frame of $\{x^{\mu}\} = \{\bar{x}^{\mu}\}$), is identified with the identity matrix $Dx^{\mu}/D\bar{x}^{\nu} = \delta^{\mu}_{\nu}$. Its difference with the coordinate transformation $\partial x^{\mu}/\partial\bar{x}^{\nu}$ must then be the Christoffel symbols as dictated by (11.27).

Just as the covariant constancy of the metric tensor is the key ingredient in the proof that the LEF exists (Box 11.1), we also have the reverse statement that the existence of an LEF proves that the metric tensor must be covariantly constant, (Problem 11.4).

11.2 Parallel transport

Parallel transport is a fundamental notion in differential geometry. It illuminates the idea of covariant differentiation, and the associated Christoffel symbols. Furthermore, using this operation, we can present another view of the geodesic as the "straightest possible curve"—geodesic line as the curve traced out by the parallel transport of its tangent vector. In Section 11.3 we shall derive the Riemann curvature tensor by way of parallel transporting a vector around a closed path.

11.2.1 Component changes under parallel transport

Equation (11.27) follows from (11.18) and expresses the relation between ordinary and covariant derivatives. To simplify the notation we write $DV^{\mu} = (D_{\nu}V^{\mu})dx^{\nu}$ and $dV^{\mu} = (\partial_{\nu}V^{\mu})dx^{\nu}$. Thus (11.18) takes on the form of

$$dV^{\mu} = DV^{\mu} - \Gamma^{\mu}_{\nu\lambda}V^{\nu}dx^{\lambda}. \tag{11.41}$$

We will show that the Christoffel symbols in the above equation reflect the effects of parallel transport of a vector by a distance of dx. First, what is a parallel transport? Why does one need to perform such a displacement? Recall the definition of the differentiation for the case of a scalar function $\Phi(x)$,

$$\frac{d\Phi(x)}{dx} = \lim_{\Delta x \to 0} \frac{\Phi(x+\Delta x) - \Phi(x)}{\Delta x}. \tag{11.42}$$

Namely, it is the difference of functional values **at two different positions**. For the coordinate-independent scalar function $\Phi(x)$, this issue of two locations does not introduce any complication. This is not the case for vector components. The differential dV^μ on the left-hand side (LHS) (11.41) is the difference

$$\begin{aligned} dV^\mu &= \lim_{\Delta x \to 0} [V^\mu(x+\Delta x) - V^\mu(x)] \\ &\equiv [V^\mu_{(2)} - V^\mu_{(1)}] \end{aligned}$$

of the vector components $V^\mu = \mathbf{e}^\mu \cdot \mathbf{V}$ evaluated at two different positions (1) and (2) separated by dx. There are two sources for their difference: the change of the vector itself, $\mathbf{V}_{(2)} \neq \mathbf{V}_{(1)}$, and a coordinate change $\mathbf{e}^\mu_{(2)} \neq \mathbf{e}^\mu_{(1)}$, corresponding to the two terms on the RHS of (11.18). Thus the total change, as given by dV^μ, is the sum of two terms

$$[\Delta V^\mu]_{\text{total}} = [\Delta V^\mu]_{\text{true}} + [\Delta V^\mu]_{\text{coord}} \tag{11.43}$$

with one term representing the change of the vector itself $\mathbf{e}^\mu \cdot d\mathbf{V} = DV^\mu$ which may be called the "true change," and another term

$$\mathbf{V} \cdot d\mathbf{e}^\mu = -\Gamma^\mu_{\nu\lambda} V^\nu dx^\lambda \tag{11.44}$$

representing the coordinate change between the two points separated by dx. The coordinate change is expected to be proportional to the vector component V^ν and to the separation dx^λ with the proportional constants in (11.44) being identified with the Christoffel symbols.

This discussion motivates us to introduce the geometric concept of **parallel transport**. It is the process of moving a vector without changing the vector itself $[\Delta V^\mu]_{\text{true}} = \mathbf{e}^\mu \cdot d\mathbf{V} = DV^\mu = 0$. As a result, the entire change of vector components under parallel displacement is due to coordinate changes. In a flat space with a Cartesian coordinate system, this is trivial as there is no coordinate change from point-to-point. But even in a flat space with a curvilinear coordinate system, such as the polar coordinates, this parallel transport itself induces

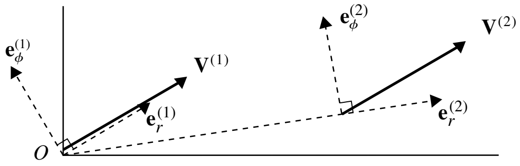

Fig. 11.1 Parallel transport of a vector $\mathbf{V}$ in a flat plane with polar coordinates: from the position-1 at the origin $\mathbf{V}^{(1)} = (V_\phi^{(1)}, V_r^{(1)})$ to another position-2, $\mathbf{V}^{(2)} = (V_\phi^{(2)}, V_r^{(2)})$. The differences of the basis vectors at these two positions $(\mathbf{e}_\phi^{(1)}, \mathbf{e}_r^{(1)}) \neq (\mathbf{e}_\phi^{(2)}, \mathbf{e}_r^{(2)})$ bring about component differences $(V_\phi^{(1)}, V_r^{(1)}) \neq (V_\phi^{(2)}, V_r^{(2)})$.

component changes. In Fig. 11.1 we have parallel transported a vector from position 1 to another position 2, $\mathbf{V}^{(1)} \to \mathbf{V}^{(2)}$. As one can see, the components have changed. In particular, $V_\phi^{(1)} = 0 \neq V_\phi^{(2)}$, and $V_r^{(1)} = \{[V_\phi^{(2)}]^2 + [V_r^{(2)}]^2\}^{1/2}$.

We can now see the connection of differentiating tensor components and parallel transport of the tensor. Differentiation always involves taking the difference of a tensor at two different positions; since tensor components are coordinate dependent, we must first parallel transport the tensor to one point, to make this comparison. The process of parallel displacement introduces changes because of coordinate changes. Thus, the total change, as represented by the ordinary differentiation, is the sum of the change of the vector itself ("truce change" as measured by the covariant differential) and of the coordinate change incurred by parallel transport, as represented by the affine connection term.

11.2.2 The geodesic as the straightest possible curve

The above discussion leads us to a mathematical expression for a "parallel transport of vector components," by setting the true change to zero in Eq. (11.41) as the vector itself does not change under such a transport:

$$DV^\mu = dV^\mu + \Gamma^\mu_{\nu\lambda} V^\nu dx^\lambda = 0. \tag{11.45}$$

Remark: Recall that we have shown the metric tensor to be covariantly constant, $D_\mu g_{\nu\lambda} = 0$. We now understand covariant constancy to mean the change of tensor components due to coordinate change only. But a change of the metric, by definition, is a pure coordinate change. Hence, it must be covariantly constant.

The process of parallel transporting a vector V^μ along a curve $x^\mu(\sigma)$ can be expressed according to (11.45) as

$$\frac{DV^\mu}{D\sigma} = \frac{dV^\mu}{d\sigma} + \Gamma^\mu_{\nu\lambda} V^\nu \frac{dx^\lambda}{d\sigma} = 0. \tag{11.46}$$

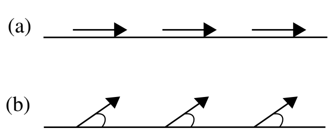

Fig. 11.2 Straight line as the geodesic in a flat plane: (a) As a curve traced out by parallel transport of its tangents. (b) When the vector is parallel transported along the straight line, the angle between them is unchanged.

From this we can define the geodesic line, as the straightest possible curve, by the condition of it being the line constructed by parallel transport of its tangent vector. See Fig. 11.2(a) for an illustration of such an operation in the flat space. In this way the condition can be formulated by setting $V^\mu = dx^\mu/d\sigma$ in (11.46):

$$\frac{D}{D\sigma}\left(\frac{dx^\mu}{d\sigma}\right) = 0. \tag{11.47}$$

Or, more explicitly,

$$\frac{d}{d\sigma}\frac{dx^\mu}{d\sigma} + \Gamma^\mu_{\nu\lambda}\frac{dx^\nu}{d\sigma}\frac{dx^\lambda}{d\sigma} = 0. \tag{11.48}$$

This agrees with the geodesic equation as shown in Chapter 5, Eq. (5.9).

Example 11.1 *When a vector V_μ is parallel transported along a geodesic, we can show that the angle subtended by the vector and the geodesic (i.e. the tangent of the geodesic) is unchanged, see Fig. 11.2(b). Namely, we need to show*

$$\frac{D}{D\sigma}\left(V_\mu \frac{dx^\mu}{d\sigma}\right) = 0. \tag{11.49}$$

The proof is straightforward:

$$\frac{D}{D\sigma}\left(V_\mu \frac{dx^\mu}{d\sigma}\right) = \frac{DV_\mu}{D\sigma}\left(\frac{dx^\mu}{d\sigma}\right) + V_\mu \frac{D}{D\sigma}\left(\frac{dx^\mu}{d\sigma}\right). \tag{11.50}$$

The RHS indeed vanishes: the first term is zero because we are parallel transporting the vector (11.46); the second term is zero because the curve is a geodesic satisfying Eq. (11.47).

11.3 Riemannian curvature tensor

Curvature measures how much a curved space is curved because it measures the amount of deviation of any geometric relations from their corresponding Euclidean equalities. We have already proven in Section 4.3.3 a particular relation showing that for a 2D curved surface the angular excess ϵ (sum of the interior angles in excess of its Euclidean value) of an infinitesimal polygon is proportional to the Gaussian curvature K at this location:

$$\epsilon = K\sigma, \tag{11.51}$$

where σ is the area of the polygon. In the following section, this relation (11.51) will be used to generalize the notion of curvature K to that of an n-dimensional curved space.

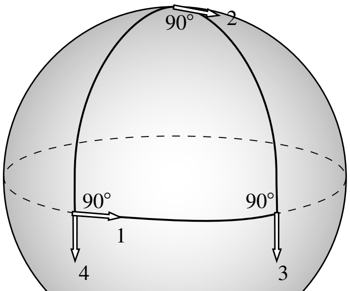

Fig. 11.3 A triangle with all interior angles being 90° on a spherical surface. The parallel transport of a vector around this triangle (clockwise from 1, to 2, to 3, and finally back to the starting point at 4) leads to a directional change of the vector by 90° (the angular difference between the vectors at point 4 and point 1).

Angular excess ϵ and directional change of a vector

How can an angular excess be "measured" in general? To implement this, we first use the concept of parallel transport to cast this relation (11.51) in a form that allows for such an n-dimensional generalization. It can be shown that angular excess ϵ is related to the directional change of a vector after it has been parallel transported around the perimeter of the polygon. The simplest example is a spherical triangle with three 90° interior angles. In Fig. 11.3 we see that a parallel transported vector changes its direction by 90°, which is the angular excess of this triangle. The generalization to an arbitrary triangle, hence to any polygon, is assigned as an exercise (Problem 11.6).

Recall the definition of an angle being the ratio of arc length to the radius, Fig. 11.4(a). Hence, the directional angular change can be written as the ratio of change of a vector to its magnitude. In this way we can relate the angular excess ϵ to the change of a vector after a transport: $\epsilon V = dV$. Substituting this into (11.51), we obtain

$$dV = KV\sigma. \tag{11.52}$$

Namely, the change of a vector after a round-trip parallel transport is proportional to the vector itself and the area of the closed path. The coefficient of proportionality is identified as the curvature.

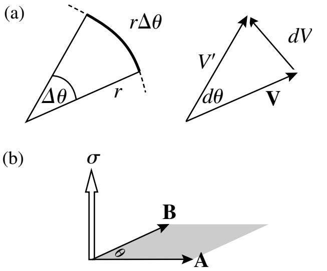

Fig. 11.4 (a) The directional change of a vector can be expressed as a fractional change of the vector: $d\theta = |d\mathbf{V}|/|\mathbf{V}|$. (b) Area of parallelogram as the cross product of its two sides, $\sigma = \mathbf{A} \times \mathbf{B}$.

The area pseudo-tensor

We will use this relation (11.52) to seek out the curvature for a higher dimensional curved space. To do so, we need to write the 2D (11.52) in the proper index form so as to be generalized to an n-dimensional space. Recall the 2D area (of a parallelogram spanned by two vectors **A** and **B**) can be calculated as a vector product, Fig. 11.4(b): $\boldsymbol{\sigma} = \mathbf{A} \times \mathbf{B}$. Or, using the antisymmetric Levi–Civita symbol in the index notation,[2]

$$\sigma_k = \epsilon_{ijk} A^i B^j. \tag{11.53}$$

Namely, $\boldsymbol{\sigma}$ has the magnitude $AB \sin\theta$ and the direction given by the right-hand-rule. But (11.53) is not a convenient form to use in higher dimension space: (i) it refers to the embedding space with a 3-valued index $i = 1, 2, 3$. (ii) For different dimensions we would need to use the antisymmetric tensor with a different number of indices, for example, for four dimensions, ϵ_{ijkl}, etc. We will instead use a two-index object σ^{ij} to represent the area:

$$\sigma^{ij} \equiv \epsilon^{ijk}\sigma_k = \epsilon^{ijk}\epsilon_{mnk}A^m B^n = \frac{1}{2}(A^i B^j - A^j B^i), \tag{11.54}$$

where we have used the identity

$$\epsilon^{ijk}\epsilon_{mnk} = \frac{1}{2}(\delta^i_m \delta^j_n - \delta^i_n \delta^j_m). \tag{11.55}$$

Furthermore, since the index $i = 3$ is irrelevant, we can write this entirely with the 2D indices $(i = 1, 2, 3) \rightarrow (a = 1, 2)$ without any reference to the embedding space. For an area in an n-dimensional space, we can represent the area by $\sigma^{\mu\nu}$ with $\mu = 1, 2, \ldots, n$

$$\sigma^{\lambda\rho} = \frac{1}{2}(A^\lambda B^\rho - B^\lambda A^\rho). \tag{11.56}$$

11.3.1 The curvature tensor in an n-dimensional space

Equation (11.52) with the area tensor of (11.56) suggests that we can represent the change dV^μ of a vector due to a parallel transport around a parallelogram spanned by A^λ and B^ρ by a tensor equation,

$$dV^\mu = R^\mu_{\ \nu\lambda\rho} V^\nu \sigma^{\lambda\rho}. \tag{11.57}$$

The change is proportional to the vector itself and to the area of the closed path. The coefficient of proportionality is a quantity with four indices and defined to be the curvature of this n-dimensional space, called the **Riemann curvature tensor**. We can plausibly expect this coefficient $R^\mu_{\nu\lambda\rho}$ to be a tensor because the differential dV^μ, being taken at the same position, is itself a good vector. With V^ν and $\sigma^{\lambda\rho}$ being tensors, the quotient theorem tells us that $R^\mu_{\nu\lambda\rho}$ should be a good tensor of rank 4 (i.e. a tensor with four indices). Explicit calculation of the parallel transport of a vector around an infinitesimal parallelogram in Box 11.2 leads to the expression:

$$R^\mu_{\ \lambda\alpha\beta} = \partial_\alpha \Gamma^\mu_{\lambda\beta} - \partial_\beta \Gamma^\mu_{\lambda\alpha} + \Gamma^\mu_{\nu\alpha}\Gamma^\nu_{\lambda\beta} - \Gamma^\mu_{\nu\beta}\Gamma^\nu_{\lambda\alpha}. \tag{11.58}$$

The Christoffel symbol Γ being first derivative, the Riemann curvature $R = d\Gamma + \Gamma\Gamma$ is then a nonlinear second derivative function of the metric $[\partial^2 g + (\partial g)^2]$.

[2] Levi–Civita symbols are discussed in sidenote 3 of Chapter 10.

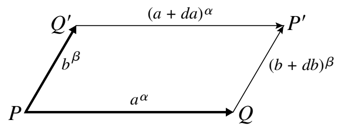

Fig. 11.5 The parallelogram $PQP'Q'$ is spanned by two vectors a^α and b^β. The opposite sides $(a+da)^\alpha$ and $(b+db)^\beta$ are obtained by parallel transport of a^α and b^β, respectively.

Box 11.2 $R^{\mu}_{\lambda\alpha\beta}$ from parallel transporting a vector around a closed path

To fix the form of the Riemann tensor as in (11.57), we carry out the operation of parallel transport for a vector around an infinitesimal parallelogram ($PQP'Q'$) spanned by two infinitesimal vectors, a^α and b^β in Fig. 11.5. Recall that parallel transport of a vector $DV^\mu = 0$ means that the total vectorial change is due entirely to coordinate change, see (11.45):

$$dV^\mu = -\Gamma^\mu_{\nu\lambda} V^\nu dx^\lambda. \tag{11.59}$$

The opposite sides of the parallelogram in Fig. 11.5, $(a+da)^\alpha$ and $(b+db)^\beta$ are obtained by parallel transport of a^α and b^β, respectively. Namely, $da^\mu = b^\mu$ and $db^\mu = a^\mu$. Applying (11.59) to these cases:

$$\begin{aligned}(a+da)^\alpha &= a^\alpha - \Gamma^\alpha_{\mu\nu} a^\mu b^\nu,\\ (b+db)^\beta &= b^\beta - \Gamma^\beta_{\mu\nu} a^\mu b^\nu.\end{aligned} \tag{11.60}$$

We now calculate, via (11.59), the change of a vector V^μ due to parallel transport from $P \to Q \to P'$:

$$\begin{aligned}dV^\mu_{PQP'} &= dV^\mu_{PQ} + dV^\mu_{QP'}\\ &= -(\Gamma^\mu_{\nu\alpha} V^\nu)_P a^\alpha - (\Gamma^\mu_{\nu\beta} V^\nu)_Q (b+db)^\beta.\end{aligned} \tag{11.61}$$

The subscripts P and Q denote the respective positions where these functions are to be evaluated. Since eventually we shall compare all quantities at one position, say P, we will Taylor expand the quantities $(\cdots)_Q$ around the point P:

$$\begin{aligned}(\Gamma^\mu_{\nu\beta})_Q &= (\Gamma^\mu_{\nu\beta})_P + a^\alpha(\partial_\alpha \Gamma^\mu_{\nu\beta})_P,\\ (V^\nu)_Q &= (V^\nu)_P + a^\alpha(\partial_\alpha V^\nu)_P = (V^\nu)_P - a^\alpha(\Gamma^\nu_{\lambda\alpha} V^\lambda)_P,\end{aligned} \tag{11.62}$$

where we have used (11.59) to reach the last expression. From now on we shall drop the subscript P. Substitute into (11.61) the expansions of (11.60) and (11.62):

$$\begin{aligned}dV^\mu_{PQP'} = &-\Gamma^\mu_{\nu\alpha} V^\nu a^\alpha - (\Gamma^\mu_{\nu\beta} + a^\alpha \partial_\alpha \Gamma^\mu_{\nu\beta})(V^\nu - a^\alpha \Gamma^\nu_{\lambda\alpha} V^\lambda)\\ &\times (b^\beta - \Gamma^\beta_{\rho\sigma} a^\rho b^\sigma).\end{aligned} \tag{11.63}$$

Multiply it out and keep terms up to $O(ab)$, we have

$$\begin{aligned}dV^\mu_{PQP'} = &-\Gamma^\mu_{\nu\alpha} V^\nu a^\alpha - \Gamma^\mu_{\nu\beta} V^\nu b^\beta + V^\nu \Gamma^\mu_{\nu\beta} \Gamma^\beta_{\rho\sigma} a^\rho b^\sigma\\ &- \partial_\alpha \Gamma^\mu_{\lambda\beta} V^\lambda a^\alpha b^\beta + \Gamma^\mu_{\nu\beta} \Gamma^\nu_{\lambda\alpha} V^\lambda a^\alpha b^\beta.\end{aligned} \tag{11.64}$$

The vectorial change due to parallel transport along the other sides: $P \to Q' \to P'$ can be obtained from the above equation simply by the interchange of $a \leftrightarrow b$.

$$\begin{aligned}dV^\mu_{PQ'P'} = &-\Gamma^\mu_{\nu\alpha} V^\nu b^\alpha - \Gamma^\mu_{\nu\beta} V^\nu a^\beta + V^\nu \Gamma^\mu_{\nu\beta} \Gamma^\beta_{\rho\sigma} a^\rho b^\sigma\\ &- \partial_\beta \Gamma^\mu_{\lambda\alpha} V^\lambda a^\alpha b^\beta + \Gamma^\mu_{\nu\alpha} \Gamma^\nu_{\lambda\beta} V^\lambda a^\alpha b^\beta.\end{aligned} \tag{11.65}$$

For a round-trip parallel transport from P back to P, the vectorial change dV^μ corresponds to the difference of the above two equations (which results in the cancellation of the first three terms on the RHS):

$$\begin{aligned} dV^{\mu} &= dV^{\mu}_{PQ'P'} - dV^{\mu}_{PQP'} \\ &= [\partial_{\alpha}\Gamma^{\mu}_{\lambda\beta} - \partial_{\beta}\Gamma^{\mu}_{\lambda\alpha} + \Gamma^{\mu}_{\nu\alpha}\Gamma^{\nu}_{\lambda\beta} - \Gamma^{\mu}_{\nu\beta}\Gamma^{\nu}_{\lambda\alpha}]V^{\lambda}a^{\alpha}b^{\beta}. \end{aligned} \tag{11.66}$$

Because the combination in the square bracket is antisymmetric with respect to the indices α and β, only the antisymmetric combination $\frac{1}{2}(a^{\alpha}b^{\beta} - a^{\beta}b^{\alpha})$ contributes. This is just the area tensor $\sigma^{\alpha\beta}$ identified in (11.56). We conclude, after a comparison of (11.66) with (11.57), that the sought-after Riemann curvature tensor in terms of Christoffel symbols is just the quoted result of (11.58).

Since this expression (11.58) for the curvature is in terms of $\Gamma^{\mu}_{\lambda\beta}$ which are not tensor components, it is not clear that $R^{\mu}_{\lambda\alpha\beta}$ has the proper tensor transformation property. However, we can show (Problem 11.7) that it can be written as the commutator of covariant derivatives:

$$[D_{\alpha}, D_{\beta}]V^{\mu} = R^{\mu}_{\ \lambda\alpha\beta}V^{\lambda}. \tag{11.67}$$

Since the covariant differentials $D_{\alpha}s$, together with $V^{\nu}s$, are good vectors—thus, according to the quotient theorem, $R^{\mu}_{\lambda\alpha\beta}$ must be a tensor. This is in agreement with our expectation, based on a heuristic argument given just prior to (11.58).

In a flat space, one can always find a coordinate system so that the metric is position-independent. Namely, not only do the first derivatives vanish, $\partial g = 0$ (as in any LEF), but also the second derivatives of the metric $\partial^2 g = 0$ likewise. In such a coordinate frame, $R^{\mu}_{\lambda\alpha\beta} \propto [\partial^2 g + (\partial g)^2] = 0$. Since it is a good tensor, if it vanishes in one set of coordinates, it vanishes for all coordinates. In fact we can also show that this is a sufficient condition for a space to be flat.

11.3.2 Symmetries and contractions of the curvature tensor

We discuss the symmetries of the Riemann curvature tensor, and counting its independent components. We note that the Riemann curvature tensor with all lower indices

$$R_{\mu\nu\alpha\beta} = g_{\mu\lambda}R^{\lambda}_{\ \nu\alpha\beta} \tag{11.68}$$

has the following symmetry features (Problem 11.8):

- It is **antisymmetric** with respect to the interchange of the first and second indices, and that of the third and fourth indices, respectively:

$$R_{\mu\nu\alpha\beta} = -R_{\nu\mu\alpha\beta}, \tag{11.69}$$

$$R_{\mu\nu\alpha\beta} = -R_{\mu\nu\beta\alpha}. \tag{11.70}$$

- It is **symmetric** with respect to the interchange of the pair made up of first and second indices with the pair of third and fourth indices:

$$R_{\mu\nu\alpha\beta} = +R_{\alpha\beta\mu\nu}. \tag{11.71}$$

- It also has the **cyclic** symmetry of

$$R_{\mu\nu\alpha\beta} + R_{\mu\beta\nu\alpha} + R_{\mu\alpha\beta\nu} = 0. \tag{11.72}$$

Knowing its symmetry properties, we can calculate the number of independent components of a curvature tensor in an n-dimensional space (Problem 11.9),

$$N_{(n)} = \frac{1}{12}n^2(n^2 - 1). \tag{11.73}$$

For various dimensions n this gives the following numbers:

- **Line**: $N_{(1)} = 0$. It is not possible for a one-dimensional inhabitant to see any curvature.
- **Surface**: $N_{(2)} = 1$. This is just the Gaussian curvature. One can check (Problem 11.11) that the expression in (11.58) corresponds to

  $$K = -\frac{R_{1212}}{\det g}, \tag{11.74}$$

  where $\det g = g_{11}g_{22} - g_{12}^2$ is the determinant of the metric tensor.
- **Spacetime**: $N_{(4)} = 20$. There are twenty independent components in the curvature tensor for our curved spacetime.
- **Metric's second derivatives**: It can be shown (Problem 11.10) that the number in (11.73) just matches that for the independent second derivatives of the metric tensor.

Contractions of the curvature tensor

Because of the symmetry properties discussed above, contractions of the curvature tensor are essentially unique. We also show how the covariantly constant **Einstein tensor**, which appears in the GR field equation (the Einstein equation) arises from contractions of the Riemann tensor.

Ricci tensor $R_{\mu\nu}$ It is the Riemann curvature tensor with the first and third indices contracted,

$$R_{\mu\nu} \equiv g^{\alpha\beta}R_{\alpha\mu\beta\nu} = R^{\beta}{}_{\mu\beta\nu}, \tag{11.75}$$

which is a symmetric tensor,

$$R_{\mu\nu} = R_{\nu\mu}. \tag{11.76}$$

It is straightforward to show that, because of the symmetry relations, the alternative contraction leads to the same Ricci tensor[3]:

$$R_{\mu\nu} = -g^{\alpha\beta}R_{\mu\alpha\beta\nu}. \tag{11.77}$$

Ricci scalar R It is the Riemann curvature tensor contracted twice,

$$R \equiv g^{\alpha\beta}R_{\alpha\beta} = R^{\beta}{}_{\beta}. \tag{11.78}$$

Bianchi identities and the Einstein tensor

There is set of constraints (called the **Bianchi identities**) on the curvature tensor:

$$D_\lambda R_{\mu\nu\alpha\beta} + D_\nu R_{\lambda\mu\alpha\beta} + D_\mu R_{\nu\lambda\alpha\beta} = 0. \tag{11.79}$$

This can be most easily proven when we go to the locally Euclidean frame (Problem 12.12). We note its resemblance to the homogeneous Maxwell equation as displayed in (10.65). There is a close analogy[4] between the curvature tensor $R_{\mu\nu\alpha\beta}$ and the electromagnetic field tensor $F_{\mu\nu}$.

[3] Note: in effect we have made a choice for sign convention in the definition of the Ricci tensor. For other sign conventions in our presentation see further comments in the next chapter when we present the GR field equation.

[4] In fact, the structure of electromagnetism can best be understood through its basic property of gauge symmetry, which has deep Riemannian geometric interpretation.

We now perform contractions on these Bianchi identities. Contracting with $g^{\mu\alpha}$ (the metric tensor being covariantly constant, $D_\lambda g^{\alpha\beta} = 0$, this metric contraction can be pushed right through the covariant differentiation):

$$D_\lambda R_{\nu\beta} - D_\nu R_{\lambda\beta} + D_\mu g^{\mu\alpha} R_{\nu\lambda\alpha\beta} = 0. \tag{11.80}$$

Contracting another time with $g^{\nu\beta}$,

$$D_\lambda R - D_\nu g^{\nu\beta} R_{\lambda\beta} - D_\mu g^{\mu\alpha} R_{\lambda\alpha} = 0. \tag{11.81}$$

At the last two terms, the metric just raises the indices,

$$D_\lambda R - D_\nu R^\nu_\lambda - D_\mu R^\mu_\lambda = D_\lambda R - 2D_\nu R^\nu_\lambda = 0. \tag{11.82}$$

Pushing through yet another $g^{\mu\lambda}$ in order to raise the λ index at the last term,

$$D_\lambda (R g^{\mu\lambda} - 2R^{\mu\lambda}) = 0. \tag{11.83}$$

Thus we see that the combination,

$$G^{\mu\nu} = R^{\mu\nu} - \frac{1}{2} R g^{\mu\nu} \tag{11.84}$$

is covariantly constant (i.e. divergence free with respect to covariant differentiation),

$$D_\mu G^{\mu\nu} = 0. \tag{11.85}$$

To summarize, $G^{\mu\nu}$, called the **Einstein tensor**, is a covariant-constant rank-2 symmetric tensor involving the second derivatives of the metric $\partial^2 g$ as well as the quadratic in ∂g.

$G^{\mu\nu}$ has the property:
conserved (covariantly constant) symmetric rank-2 tensor $\partial\Gamma, \Gamma^2 \backsim (\partial^2 g), (\partial g)^2$.

(11.86)

As we shall see in the next chapter, this is just the sought-after mathematical quantity in the field equation of GR.

Review questions

1. Writing the coordinate transformation as a partial derivative matrix, give the transform law for a contravariant vector $V^\mu \rightarrow V'^\mu$, as well as that for a mixed tensor $T^\mu_\nu \rightarrow T'^\mu_\nu$.

2. What is the fundamental difference between the coordinate transformations in a curved space and those in flat space (e.g. Lorentz transformations in the flat Minkowski space)?

3. Given the transformation of vector components as
$$V_\mu \rightarrow V'_\mu = \frac{\partial x^\lambda}{\partial x'^\mu} V_\lambda,$$
how do the derivatives $\partial_\mu V_\nu$ change under the general coordinate transformation? How do the covariant derivatives $D_\mu V_\nu$ transform? Why is it important to have differentiations that result in tensors?

4. What is the basic reason why $\partial_\mu V_\nu$ is not a tensor?

5. Write out the covariant derivative of a general tensor $D_\mu T^{\lambda\rho}_{\nu}$ (in terms of the connection symbols).

6. The relation between Christoffel symbols and the metric tensor is called "the fundamental theorem of Riemannian geometry." Write out this relation.

7. As the Christoffel symbol $\Gamma^{\mu}_{\alpha\beta}$ is not a tensor, how do we know $R^{\mu}_{\ \lambda\alpha\beta} = \partial_\alpha \Gamma^{\mu}_{\lambda\beta} - \partial_\beta \Gamma^{\mu}_{\lambda\alpha} + \Gamma^{\mu}_{\nu\alpha}\Gamma^{\nu}_{\lambda\beta} - \Gamma^{\mu}_{\nu\beta}\Gamma^{\nu}_{\lambda\alpha}$ is really a tensor?

8. What are the two basic properties of the Einstein tensor $G^{\mu\nu} = R^{\mu\nu} - \frac{1}{2}Rg^{\mu\nu}$?

9. What is "the flatness theorem"? Use this theorem to show that the metric tensor is covariantly constant, $D_\mu g_{\nu\lambda} = 0$.

Problems

(11.1) **Covariant derivative for a covariant vector** Given that the covariant derivative for a contravariant vector has the form of Eq. (11.27), show that the covariant derivative for the covariant vector is $D_\nu V_\mu = \partial_\nu V_\mu - \Gamma^{\lambda}_{\nu\mu} V_\lambda$. (**Hint**: $V_\mu V^\mu$ is an invariant.)

(11.2) **Moving bases and Christoffel symbols in polar coordinates for a flat plane** Even in a flat space, one can have moving bases. Recall the example of the polar coordinates (r, θ) on a plane surface.

(a) Work out their respective (moving) base vectors $(\mathbf{e}_r, \mathbf{e}_\theta)$ and $(\mathbf{e}^r, \mathbf{e}^\theta)$ in terms of the (fixed) Cartesian bases $(\mathbf{i}, \mathbf{j})$.
(b) Calculate the Christoffel symbols through their definition of $\partial_\nu \mathbf{e}^\mu = -\Gamma^{\mu}_{\nu\lambda}\mathbf{e}^\lambda$ given in (11.26).
(c) Calculate the divergence in a polar coordinate system: work out $D_\mu V^\mu = \partial_\mu V^\mu + \Gamma^{\mu}_{\mu\nu} V^\nu$ in terms of component fields (V^r, V^θ).
(d) Calculate the Laplacian $D_\mu D^\mu \Phi(x)$ in a polar coordinate system.
(e) Use the Christoffel symbols obtained in (b) to show that the metric tensors are constant with respect to covariant differentiation.
(f) Use the fundamental theorem of Riemannian geometry (11.37) to calculate a few $\Gamma^{\mu}_{\nu\lambda}$ to check the results obtained in (b).
(g) Use the explicit form of the Christoffel symbols calculated in (b) to show that the only independent component for the curvature tensor vanishes, $R_{1212} = 0$, as expected for a flat space.

(11.3) **Symmetry property of Christoffel symbols** Prove $\Gamma^{\mu}_{\nu\lambda} = \Gamma^{\mu}_{\lambda\nu}$ by an explicit computation of the double covariant derivatives of a scalar function $\Phi(x)$. (**Hint**: $D_\mu \Phi = \partial_\mu \Phi$ because $\Phi(x)$ is coordinate-independent.)

(11.4) **Metric is covariantly constant: further proofs** Besides the proof given in Eqs (11.33) and (11.34), prove $D_\lambda g_{\mu\nu} = 0$ in other ways by using:

(a) the fundamental theorem of Riemannian geometry (11.37);
(b) the existence of LEF (with the definite properties of $g_{\mu\nu}$ and $\Gamma^{\mu}_{\nu\lambda}$ in such a frame).

(11.5) **$D_\nu V_\mu$ is a good tensor: another proof** Use Eq. (11.49) and the geodesic Eq. (11.48) to prove that

$$(D_\nu V_\lambda)\frac{dx^\nu}{d\sigma}\frac{dx^\lambda}{d\sigma} = 0.$$

This is another way to see, via the quotient theorem, that $D_\nu V_\mu$ is a good tensor.

(11.6) **Parallel transport of a vector around a general spherical triangle** Prove that the directional change of a vector, after being parallelly transported around the perimeter of an arbitrary triangle, on a spherical surface, is equal to the angular excess of the triangle. This result then holds for any spherical polygon, since any polygon can always be divided into triangles. This in turn implies that such a relation is valid for any infinitesimal closed geodesic path in a general 2D space.

(11.7) **Riemann curvature tensor as the commutator of covariant derivatives** To show that $R^{\mu}_{\lambda\alpha\beta}$ is indeed a tensor, we can perform the following calculation: take the double derivative $D_\alpha D_\beta V^\mu = D_\alpha(\partial_\beta V^\mu + \Gamma^{\mu}_{\beta\lambda}V^\lambda) = \cdots$ as well as that in the reverse order $D_\beta D_\alpha V^\mu = D_\beta(\partial_\alpha V^\mu + \Gamma^{\mu}_{\alpha\lambda}V^\lambda) = \cdots$. Show that their difference is just the expression for the Riemann tensor as given by Eq. (11.58):

$$[D_\alpha, D_\beta]V^\mu = R^{\mu}_{\lambda\alpha\beta}V^\lambda. \qquad (11.87)$$

(11.8) **Symmetries of $R_{\mu\nu\alpha\beta}$** Since the symmetry properties are not changed by coordinate transformations, one can choose a particular coordinate frame to prove these symmetry relations, and once proven in one frame, we can then claim their validity in all frames. An obvious choice is the locally Euclidean frame with $(\Gamma = 0, \partial\Gamma \neq 0)$ where the curvature takes on a simpler form, $R_{\mu\nu\alpha\beta} = g_{\mu\lambda}(\partial_\alpha \Gamma^\lambda_{\nu\beta} - \partial_\beta \Gamma^\lambda_{\nu\alpha})$, and the symmetry properties are easy to inspect. In this way, check the validity of the symmetry properties as shown in Eqs (11.69)–(11.72).

(11.9) **Counting independent elements of Riemann tensor** The Riemann curvature tensor has the symmetry properties shown in (11.69)–(11.72). Show that the number of independent components of a curvature tensor in an n-dimensional space is $N_{(n)} = \frac{1}{12}n^2(n^2-1)$.

(11.10) **The number of metric's independent second derivatives and Riemann tensor**

(a) Calculate $A_{(n)}$, the number of independent elements in $g_{\mu\nu}$, $\partial_\alpha g_{\mu\nu}$ and $\partial_\alpha\partial_\beta g_{\mu\nu}$, taking into consideration only the symmetry properties of these tensors. First give the result $A_{(4)}$ in a 4D space, and then record the number for $\partial_\alpha\partial_\beta g_{\mu\nu}$ in a general n-dimensional space.

(b) The number $A_{(n)}$ obtained in (a) for the independent elements in $g_{\mu\nu}$, $\partial_\alpha g_{\mu\nu}$, and $\partial_\alpha\partial_\beta g_{\mu\nu}$ is an overcount, in the sense that some of them can be eliminated by coordinate transformations. If we are interested in the number of "truly independent elements" that reflects the property of the space itself (rather than the coordinate system), we should subtract out the elements that can be transformed away. Now count $B_{(4)}$, the number of elements that can be transformed away. (**Suggestion**: the transformation of the metric $g_{\mu\nu}$ is given in (10.13). The relevant transformation matrix can be written as a shorthand:

$$\frac{\partial x^\beta}{\partial x'^\alpha} \equiv (\partial_\alpha x_\beta). \qquad (11.88)$$

From our proof (in Box 11.1) of the flatness theorem by way of power series expansions, we see that transformations of tensor derivatives depend on the derivatives of the transformation matrices. For example, the first derivative $\partial_\alpha g_{\mu\nu}$ transformation is determined by the first derivative of the number of transformation $\partial_\gamma(\partial_\alpha x_\beta)$, and $\partial_\alpha\partial_\beta g_{\mu\nu}$ by $\partial_\gamma\partial_\delta(\partial_\alpha x_\beta)$, etc. The number of parameters in these transformations (and their derivatives) should be the number of elements that can be transformed away by coordinate transformations.)

(c) The difference $N_{(4)} = A_{(4)} - B_{(4)}$ obtained in (a) and (b) should correspond to the number of independent elements in $g_{\mu\nu}$, $\partial_\alpha g_{\mu\nu}$, and $\partial_\alpha\partial_\beta g_{\mu\nu}$. Do these counts make physical sense? Give your interpretation for each case.

(d) Write out $N_{(n)}$ for $\partial_\alpha\partial_\beta g_{\mu\nu}$ in the n-dimensional space. You should find a result that matches Problem 11.9.

(11.11) **Reducing Riemann tensor to Gaussian curvature** For a 2D space, the curvature tensor has only one independent element. Show that it is just the Gaussian curvature of (4.35) with the identification of

$$K = -\frac{R_{1212}}{\det g}.$$

(11.12) **Bianchi identities** Demonstrate the validity of the Bianchi identity (11.79) in the local inertial frame.

(11.13) **Ricci tensor is symmetric** From the definition of $R_{\mu\nu} \equiv g^{\alpha\beta}R_{\alpha\mu\beta\nu}$, show that $R_{\mu\nu} = R_{\nu\mu}$.

(11.14) **Contraction of Christoffel symbols** Show that

$$\Gamma^\mu_{\mu\alpha} = \frac{1}{\sqrt{-g}}\frac{\partial}{\partial x^\alpha}\sqrt{-g},$$

where g is the determinant of the matrix $g_{\mu\nu}$.

(11.15) **Contraction of Riemann tensor** Show that $R^\mu_{\ \mu\alpha\beta} = 0$. (**Hint**: use the relation obtained in Problem 11.14.) Recall that the Ricci tensor is obtained by contracting the first and third indices of the Riemann tensor. This result shows that all contractions of the Riemann tensor, based on its symmetry properties, are either the same as the Ricci tensor or are zero.

GR as a geometric theory of gravity - II

12

- The mathematical realization of the equivalence principle (EP) is the principle of general covariance. General relativity (GR) equations must be covariant with respect to general coordinate transformations.
- To go from special relativity (SR) to GR equations, one replaces ordinary by covariant derivatives. The SR equation of motion $d^2x^\mu/d\tau^2 = 0$ turns into $D^2x^\mu/D\tau^2 = 0$, which is the geodesic equation.
- The Einstein equation, as the relativistic gravitation field equation, relates the energy–momentum tensor to the Einstein curvature tensor.
- The Schwarzschild metric is shown to be the solution of the Einstein equation for the case of a spherical source.
- The solutions of Einstein's equation that satisfy the cosmological principle must have a space with constant curvature—the Robertson–Walker spacetime.
- The relation of the cosmological Friedmann equations to the Einstein field equation is explicated.
- The mathematical compatibility of the cosmological constant term with Einstein equation's structure, and its interpretation as the vacuum energy tensor, are discussed.

In Chapter 5 we have presented arguments for a geometric theory of gravity. The gravitational field is identified with the warped spacetime described by the metric function $g_{\mu\nu}(x)$. After one accepts that spacetime can be curved and the Riemannian geometry as the appropriate mathematics to describe such a space, we can now use the tensor calculus learned in Chapters 10 and 11 to write down the physics equations satisfying the principle of general relativity (GR). In Section 12.1 we present the principle of general covariance, which guides us to GR equations in a curved spacetime. A proper derivation of the geodesic equation as the GR equation of motion will be presented, and we can finally write down the GR field equation, the Einstein equation. Its connection to the Newton/Poisson equation is discussed. Finally, we show how to obtain the Schwarzschild metric as the solution to the Einstein equation with a spherical source. In Section 12.4, the geometric formalisms used in cosmology as discussed in Chapters 7–9 are studied as solutions of Einstein equation compatible with the cosmological principle.

12.1 The principle of general covariance

According to the strong principle of equivalence, gravity can always be transformed away locally. The physical laws, or the field equation for $g_{\mu\nu}(x)$, must have the same form no matter what generalized coordinates are used to locate or label worldpoints (events) in spacetime. One expresses this by the requirement that the physics equations must satisfy the **principle of general covariance**. This is a two-part statement:

1. Physics equations must be covariant under the general coordinate transformations which leave the infinitesimal spacetime separation ds^2 invariant.
2. Physics equations should reduce to the correct special relativistic form in the local inertial frames. Namely, we must have the correct SR equations in the free fall frames, in which gravity is transformed away. Additionally, gravitational equations reduce to Newtonian equations in the limit of low velocity particles in a weak and static field.

This provides us with a well-defined path to go from SR equations, valid in the local inertial frames with no gravity, to GR equations that are valid in every coordinate system in the curved spacetime—curved because of the presence of gravity. Such GR equations must be covariant under general local transformations. The key feature of a general covariance transformation, in contrast to the (Lorentz) transformation in a flat spacetime, is its spacetime-dependence. The tensor formalism in a curved spacetime differs from that for a flat spacetime of SR in its derivatives. To go from an SR equation to the corresponding GR equation is simple: we need to replace the ordinary derivatives $[\partial]$ in SR equations by covariant derivatives $[D]$:

$$\partial \longrightarrow D = \partial + \Gamma. \tag{12.1}$$

Since Christoffel symbols Γ are the derivatives of the metric—hence the derivatives of the gravitational potential (i.e. the gravitational field), the introduction of covariant derivatives naturally brings the gravitational field into the physics equations. In this way we can, for example, find the equations that describe electromagnetism in the presence of a gravitational field. In Table 12.1, we show how GR equations arise from the SR results.

This discussion of introducing gravitational coupling in GR illustrates how a local symmetry can dictate the form of dynamics—in this case, the precise way the Christoffel symbol $\Gamma^{\lambda}_{\mu u}$ (gravitational field) enters into physics equations. For example, in the last line of Table 12.1, starting from the familiar special relativistic Eq. (10.62), we have the set of GR equations in curved spacetime,

$$\partial_\mu F^{\mu\nu} + \Gamma^{\mu}_{\mu\lambda} F^{\lambda\nu} + \Gamma^{\nu}_{\mu\lambda} F^{\mu\lambda} = -\frac{1}{c}\, j^{\nu}, \tag{12.2}$$

Table 12.1 SR electromagnetic equations in the flat spacetime vs. GR equations in a curved spacetime

SR equations			GR equations
Lorentz force law in flat spacetime			in curved spacetime
Eq. (10.60)	$\dfrac{dU^\mu}{d\tau} = \dfrac{q}{c} F^{\mu\nu} U_\nu$	$\longrightarrow$	$\dfrac{DU^\mu}{D\tau} = \dfrac{q}{c} F^{\mu\nu} U_\nu$
Maxwell's equation in flat spacetime			in curved spacetime
Eq. (10.66)	$\partial_\mu \tilde{F}^{\mu\nu} = 0$	$\longrightarrow$	$D_\mu \tilde{F}^{\mu\nu} = 0$
Eq. (10.62)	$\partial_\mu F^{\mu\nu} = -\dfrac{1}{c}\, j^\nu$	$\longrightarrow$	$D_\mu F^{\mu\nu} = -\dfrac{1}{c}\, j^\nu$

which are interpreted as Gauss's and Ampere's laws in the presence of a gravitational field.

12.1.1 Geodesic equation from SR equation of motion

Now that we have the GR equations for electromagnetism, what about the GR equations for gravitation? The procedure illustrated in the above case fails because there is no SR equation for the gravitational field. Thus, for gravitational field equations we need a fresh start—for guidance we need to go back to the Newtonian theory of gravitation, cf. Section 3.1.

In Table 12.1, we have already written down the gravitational equation of motion: the equation that allows us to find the motion of a test charge in the presence of electromagnetic, as well as gravitational, fields. Just concentrating on the gravitational part, hence setting the EM field tensor $F_{\mu\nu} = 0$, we have the equation of motion for a particle in a gravitational field:

$$\frac{DU^\mu}{D\tau} = 0, \tag{12.3}$$

where U^μ is the 4-velocity of the test particle, and τ is the proper time. In fact, we should think its derivation more directly as the generalization from the special relativistic equation of motion for a free particle:

$$\frac{dU^\mu}{d\tau} = 0, \tag{12.4}$$

which simply states that in the absence of an external force the test particle follows a trajectory of constant velocity.

We now demonstrate that this Eq. (12.3) is just the geodesic Eq. (5.9). Using the explicit form of the covariant differentiation (11.27), the above equation can be written as

$$\frac{DU^\mu}{D\tau} = \frac{dU^\mu}{d\tau} + \Gamma^\mu_{\nu\lambda} U^\nu \frac{dx^\lambda}{d\tau} = 0. \tag{12.5}$$

Plug in the expression of 4-velocity in terms of the position vector[1]

$$U^\mu = \frac{dx^\mu}{d\tau}, \tag{12.6}$$

we immediately obtain an equation

$$\frac{d^2x^\mu}{d\tau^2} + \Gamma^\mu_{\nu\lambda} \frac{dx^\nu}{d\tau} \frac{dx^\lambda}{d\tau} = 0, \tag{12.7}$$

which is recognized as the geodesic equation (5.9). This supports our heuristic argument—"particles should follow the shortest and straightest possible trajectories"—used in Section 5.2 to suggest that the GR equation of motion should be the geodesic equation.

[1] For the 4-velocity, we have $U^\mu = Dx^\mu/D\tau = dx^\mu/d\tau$ because $dx^\mu/d\tau$ is already a "good vector" as can been seen from the fact that $(ds/d\tau)^2 = g_{\mu\nu}(dx^\mu/d\tau)(dx^\nu/d\tau)$ is a scalar.

12.2 Einstein field equation

The equation of motion of the Newtonian theory is generalized to be the geodesic equation:

$$\left[\frac{d^2\mathbf{r}}{dt^2} = -\boldsymbol{\nabla}\Phi\right] \rightarrow \left[\frac{d^2x^\mu}{d\tau^2} + \Gamma^\mu_{\nu\lambda}\frac{dx^\nu}{d\tau}\frac{dx^\lambda}{d\tau} = 0\right]. \qquad (12.8)$$

The next step is to generalize its field equation:

$$[\nabla^2\Phi = 4\pi G_{\mathrm{N}}\rho] \rightarrow [?], \qquad (12.9)$$

where G_{N} is Newton's constant, and ρ is the mass density function, cf. Sections 3.1 and 5.1.

12.2.1 Finding the relativistic gravitational field equation

We have already learned that the metric tensor is the relativistic generalization of the gravitational potential (Section 5.1) and mass density is the (0, 0) component of the energy–momentum tensor (Section 10.3):

$$\left(1 + \frac{2\Phi(x)}{c^2}\right) \rightarrow g_{00}(x) \quad \text{and} \quad \rho(x) \rightarrow T_{00}(x). \qquad (12.10)$$

The GR field equation, being the relativistic generalization of the Newtonian field Eq. (12.9), should have the structure, when written out in operator form,

$$[\hat{O}g] = \kappa[T]. \qquad (12.11)$$

Namely, some differential operator $[\hat{O}]$ acting on the metric $[g]$ to yield the energy–momentum tensor $[T]$ with κ being the "conversion factor" proportional to Newton's constant G_{N} that allows us to relate energy density and the spacetime curvature. Since we expect $[\hat{O}g]$ to have the Newtonian limit of $\nabla^2\Phi$, $[\hat{O}]$ must be a second-order differential operator. Besides the $\partial^2 g$ terms, we also expect it to contain nonlinear operators of the type of $(\partial g)^2$. The presence of the nonlinear terms $(\partial g)^2$ is suggested by the fact that energy, just like mass, is a source of gravitational fields, and gravitational fields themselves hold energy—just as electromagnetic fields hold energy, with density being quadratic in fields $(\mathbf{E}^2+\mathbf{B}^2)$. Namely, gravitational field energy density must be quadratic in the gravitational field strength, $(\partial g)^2$. In terms of Christoffel symbols $\Gamma \sim \partial g$, we anticipate $[\hat{O}g]$ to contain not only $\partial\Gamma$ but also Γ^2 terms as well. Furthermore, because the right-hand side (RHS) is a symmetric tensor of rank 2 which is covariantly constant, $D_\mu T^{\mu\nu} = 0$ (reflecting energy–momentum conservation), the left-hand side (LHS) $[\hat{O}g]$ must have these properties also. The basic properties that the LHS of the field equation must have are summarized below:

$\hat{O}g$ must have the property:
conserved (covariantly constant) symmetric rank-2 tensor $(\partial^2 g), (\partial g)^2 \backsim \partial\Gamma, \Gamma^2$.

(12.12)

There is only one such second rank tensor: the Einstein tensor $G_{\mu\nu}$, see Section 11.3.2, Eq. (11.86). Thus Einstein proposed the GR field equation to be

$$G_{\mu\nu} = \kappa T_{\mu\nu}, \tag{12.13}$$

where the proportional constant κ will be determined when we compare this field equation with that in the Newtonian theory. Writing out the Einstein equation in terms of the Ricci scalar and tensor we have

$$R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} = \kappa T_{\mu\nu}. \tag{12.14}$$

This equation can be written in an alternative form by taking the trace of the above equation:

$$-R = \kappa T, \tag{12.15}$$

where T is the trace of the energy–momentum tensor, $T = g^{\mu\nu}T_{\mu\nu}$. In this way we can rewrite the field equation in an equivalent form by replacing $Rg_{\mu\nu}$ by $-\kappa Tg_{\mu\nu}$:

$$R_{\mu\nu} = \kappa\left(T_{\mu\nu} - \frac{1}{2}Tg_{\mu\nu}\right). \tag{12.16}$$

12.2.2 Newtonian limit of the Einstein equation

Here we shall show that the familiar Newtonian field Eq. (12.9) is simply the leading approximation to the Einstein Eq. (12.16) in the Newtonian limit defined in Section 5.2.1 as being for a nonrelativistic source particle producing a weak and static gravitational field.

Nonrelativistic velocity. In the nonrelativistic regime of small v/c, the rest energy density term T_{00} being dominant, we shall concentrate on the $(0, 0)$ component of (12.16), as other terms are down by $O(v/c)$:

$$R_{00} = k\left(T_{00} - \frac{1}{2}Tg_{00}\right) \tag{12.17}$$

with

$$T = g^{\mu\nu}T_{\mu\nu} \backsimeq g^{00}T_{00} = \frac{1}{g_{00}}T^{00}. \tag{12.18}$$

Thus (12.17) becomes

$$R_{00} = \frac{1}{2}\kappa T_{00}. \tag{12.19}$$

To recover the Newtonian field equation, we need to show that $R_{00} \to \nabla^2 g_{00}$: from the definition of Ricci tensor (in terms of the Riemann curvature tensor), we have

$$R_{00} = g^{\mu\nu}R_{\mu 0\nu 0} = g^{ij}R_{i0j0}, \tag{12.20}$$

where $(i = 1, 2, 3)$ and in reaching the last equality we have used the fact that the tensor components such as R_{0000} and R_{i000} all vanish because of symmetry properties of the curvature tensor.

Weak field limit. The Newtonian limit also corresponds to weak field limit, hence we will keep as few powers of the metric tensor as possible: that is, keep $\partial\partial g$ terms rather than $(\partial g)^2$s, etc.

$$R_{\mu\nu\alpha\beta} = \frac{1}{2}(\partial_\mu\partial_\alpha g_{\nu\beta} - \partial_\nu\partial_\alpha g_{\mu\beta} + \partial_\nu\partial_\beta g_{\mu\alpha} - \partial_\mu\partial_\beta g_{\nu\alpha}). \tag{12.21}$$

Substituting this into (12.20) we have

$$R_{00} = g^{ij}R_{i0\,j0} = \frac{g^{ij}}{2}(\partial_i\partial_j g_{00} - \partial_0\partial_j g_{i0} + \partial_{i\mu}\partial_0 g_{0j} - \partial_0\partial_{0\beta} g_{ij}). \tag{12.22}$$

Static limit. The Newtonian limit also corresponds to a static situation; we can drop in (12.22) all terms having a time derivative ∂_0 factor,

$$R_{00} = \frac{1}{2}\,\nabla^2\,g_{00}.$$

After using the relation in (5.20) and $T_{00} = \rho c^2$, Eq. (12.19) becomes

$$-\frac{1}{2}\,\nabla^2\left(1 + 2\frac{\Phi}{c^2}\right) = \frac{1}{2}\kappa\rho c^2 \tag{12.23}$$

or

$$\nabla^2\Phi = -\frac{1}{2}\kappa\rho c^4. \tag{12.24}$$

Thus we see that the Einstein equation indeed has the correct Newtonian limit of $\nabla^2\phi = 4\pi G_{\mathrm{N}}\rho$ when we identify

$$\kappa = -\frac{8\pi G_{\mathrm{N}}}{c^4}. \tag{12.25}$$

The Einstein equation

Putting this value of (12.25) into the field Eq. (12.14) we have the **Einstein equation**[2]

$$R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} = -\frac{8\pi G_{\mathrm{N}}}{c^4}T_{\mu\nu}, \tag{12.26}$$

or, written in its equivalent form as (12.16):

$$R_{\mu\nu} = -\frac{8\pi G_{\mathrm{N}}}{c^4}\left(T_{\mu\nu} - \frac{1}{2}Tg_{\mu\nu}\right). \tag{12.27}$$

This is a set of 10 coupled nonlinear partial differential equations. In general they are extremely difficult to solve. However, for the spherically symmetric (in the three spatial dimensions) situation, an analytic solution can be obtained. We shall study this solution in the following section.

Box 12.1 Newtonian limit for the general source having mass density and pressure

In certain situations, with cosmology being the notable example, we consider the source of gravity being a plasma having mass density and pressure. Usually we can drop the pressure term, which is negligible for nonrelativistic matter. However, if the corresponding matter density is particularly low, or just comparable to the pressure contribution, we need to work out the Newtonian limit for a general source with an energy–momentum tensor

[2] Beware of various sign conventions $[S] = \pm 1$ used in the literature:

$$\eta_{\mu\nu} = [S1] \times \mathrm{diag}(-1,1,1,1),$$
$$R^{\mu}_{\lambda\alpha\beta} = [S2] \times (\partial_\alpha\Gamma^{\mu}_{\lambda\beta} - \partial_\beta\Gamma^{\mu}_{\lambda\alpha},$$
$$+\,\Gamma^{\mu}_{\nu\alpha}\Gamma^{\nu}_{\lambda\beta} - \Gamma^{\mu}_{\nu\beta}\Gamma^{\nu}_{\lambda\alpha})$$
$$G_{\mu\nu} = [S3] \times \frac{8\pi G}{c^4}T_{\mu\nu}.$$

Thus our convention is $[S1, S2, S3] = (+\,+\,-)$. The sign in the Einstein equation $[S3]$ is related to the sign convention in the definition of the Ricci tensor $R_{\mu\nu} = R^{\alpha}_{\mu\alpha\nu}$.

of the ideal fluid as in (10.88). As shown in (12.17), the dominant term in this limit is the 00-component of the Einstein Eq. (12.16),

$$\begin{aligned} R_{00} &= \kappa \left(T_{00} - \frac{1}{2} T g_{00} \right) = \frac{\kappa}{2} (T_{00} + T_{11} + T_{22} + T_{33}) \\ &= \frac{\kappa}{2} (\rho c^2 + 3p). \end{aligned} \tag{12.28}$$

The trace (T) of the energy–momentum tensor, to the leading order in the Newtonian limit, has been calculated by using the flat spacetime metric: $T = \eta^{\mu\nu} T_{\mu\nu}$ with $\eta^{\mu\nu} = \text{diag}(-1, 1, 1, 1)$. From this we obtain the quasi-Newtonian equation for the gravitational potential Φ, as first displayed in (9.4):

$$\nabla^2 \Phi = 4\pi G_{\text{N}} \left(\rho + 3\frac{p}{c^2} \right). \tag{12.29}$$

This makes it clear that not only mass, but also pressure, can be a source of gravitational field.

12.3 The Schwarzschild exterior solution

We now solve the Einstein equation for a spherically symmetric (nonrotating) source with total mass M. The solution is the metric function $g_{\mu\nu}(x)$ for the spacetime geometry outside the source, and is called the Schwarzschild exterior solution. In Section 6.1.1 we have shown that a spatially spherical symmetric metric tensor (6.12) has only two scalar unknown functions:

$$ds^2 = g_{00}(r,t) c^2 dt^2 + g_{rr}(r,t) dr^2 + r^2 (d\theta^2 + \sin^2\theta d\phi^2). \tag{12.30}$$

Here we shall use the Einstein equation to solve for g_{00} and g_{rr}. The first step involves expressing the Ricci tensor elements $R_{\mu\nu}$ in terms of these metric elements.

The spherical symmetric Christoffel symbols

We begin by calculating the connection symbols based on the spherically symmetric form of (12.30). It will be convenient to introduce the notation:

$$g_{00} = \frac{1}{g^{00}} \equiv -e^{\nu}, \qquad g_{rr} = \frac{1}{g^{rr}} \equiv e^{\rho} \tag{12.31}$$

so that the unknown metric functions are now $\nu(r,t)$ and $\rho(r,t)$. Here we state the result (see Box 12.2 for comments on the calculational procedure):

$$\begin{aligned} &\Gamma^0_{00} = -\frac{\dot{\nu}}{2}, \quad \Gamma^0_{rr} = \frac{\dot{\rho}}{2} e^{\rho-\nu}, \quad \Gamma^0_{0r} = \frac{\nu'}{2}, \\ &\Gamma^r_{00} = \frac{\nu'}{2} e^{\nu-\rho}, \quad \Gamma^r_{rr} = \frac{\rho'}{2}, \quad \Gamma^r_{0r} = \frac{\dot{\rho}}{2}, \\ &\Gamma^r_{\theta\theta} = -re^{-\rho}, \quad \Gamma^r_{\phi\phi} = -r\sin^2\theta e^{-\rho}, \quad \Gamma^\theta_{\phi\phi} = -\sin\theta\cos\theta, \\ &\Gamma^\theta_{r\theta} = r^{-1}, \quad \Gamma^\phi_{\phi\theta} = \cot\theta, \quad \Gamma^\phi_{r\phi} = r^{-1}, \end{aligned} \tag{12.32}$$

where dot denotes differentiation with respect to the coordinate time $x^0 = ct$, while the prime is a differentiation with respect to the radial coordinate r: for example,

$$\dot{\nu} = \frac{1}{c}\frac{\partial \nu}{\partial t}, \qquad \nu' = \frac{\partial \nu}{\partial r}. \tag{12.33}$$

Box 12.2 $\Gamma^{\mu}_{\nu\lambda}$ via the Euler–Lagrange equation

In principle, we can obtain the result in (12.32) by differentiating the metric tensor as in (11.37). A more efficient procedure will be through the interpretation of the geodesic equation

$$\frac{d^2x^\mu}{d\sigma^2} + \Gamma^{\mu}_{\nu\lambda}\frac{dx^\nu}{d\sigma}\frac{dx^\lambda}{d\sigma} = 0 \tag{12.34}$$

as the Euler–Lagrange equation

$$\frac{d}{d\sigma}\frac{\partial L}{\partial \dot{x}^\mu} - \frac{\partial L}{\partial x^\mu} = 0 \tag{12.35}$$

with the Lagrangian being (see Section 4.2.1 for more detail)

$$L = \frac{1}{2}g_{\mu\nu}\frac{dx^\mu}{d\sigma}\frac{dx^\nu}{d\sigma} = \frac{1}{2}\left[-e^\nu\left(\frac{dx^0}{d\sigma}\right)^2 + e^\rho\left(\frac{dr}{d\sigma}\right)^2 + r^2\left(\frac{d\theta}{d\sigma}\right)^2 + r^2\sin^2\theta\left(\frac{d\phi}{d\sigma}\right)^2\right]. \tag{12.36}$$

Once the geodesic equation is written out this way as in (12.35), we can then extract the value of $\Gamma^{\mu}_{\nu\lambda}$ by comparing it to (12.34). For example, because we have

$$\frac{\partial L}{\partial x^0} = \frac{1}{2}\left[-\dot{\nu}e^\nu\left(\frac{dx^0}{d\sigma}\right)^2 + \dot{\rho}e^\rho\left(\frac{dr}{d\sigma}\right)^2\right] \quad \text{and} \quad \frac{\partial L}{\partial \dot{x}^0} = -e^\nu\left(\frac{dx^0}{d\sigma}\right)$$

the $\mu = 0$ component of the Euler–Lagrange Eq. (12.35) reads as

$$\frac{d}{d\sigma}\left[-e^\nu\left(\frac{dx^0}{d\sigma}\right)\right] - \frac{1}{2}\left[-\dot{\nu}e^\nu\left(\frac{dx^0}{d\sigma}\right)^2 + \dot{\rho}e^\rho\left(\frac{dr}{d\sigma}\right)^2\right] = 0$$

or

$$-e^\nu\left[\frac{d^2x^0}{d\sigma^2} + \nu'\frac{dr}{d\sigma}\frac{dx^0}{d\sigma} - \frac{\dot{\nu}}{2}\left(\frac{dx^0}{d\sigma}\right)^2 + \frac{\dot{\rho}}{2}e^{\rho-\nu}\left(\frac{dr}{d\sigma}\right)^2\right] = 0.$$

This is to be compared to the $\mu = 0$ component of (12.34), which with only the nonvanishing $(dx^\nu/d\sigma)(dx^\lambda/d\sigma)$ factors displayed, has the form:

$$\frac{d^2x^0}{d\sigma^2} + 2\Gamma^0_{r0}\frac{dr}{d\sigma}\frac{dx^0}{d\sigma} + \Gamma^0_{00}\left(\frac{dx^0}{d\sigma}\right)^2 + \Gamma^0_{rr}\left(\frac{dr}{d\sigma}\right)^2 = 0.$$

Hence we can extract the result:

$$\Gamma^0_{r0} = \frac{\nu'}{2}, \quad \Gamma^0_{00} = -\frac{\dot{\nu}}{2}, \quad \Gamma^0_{rr} = \frac{\dot{\rho}}{2}e^{\rho-\nu}, \tag{12.37}$$

as displayed in (12.32).

The spherically symmetric curvature

From the Christoffel symbols we then use (11.58) to calculate the curvature tensor $R^{\alpha}_{\mu\beta\nu}$ from which we can contract the indices $R^{\alpha}_{\mu\alpha\nu}$ to form the Ricci tensor:

$$\begin{aligned}
R_{00} &= -\left(\frac{\nu''}{2}+\frac{\nu'^2}{4}-\frac{\nu'\rho'}{4}+\frac{\nu'}{r}\right)e^{\nu-\rho}+\left(\frac{\ddot{\rho}}{2}+\frac{\dot{\rho}^2}{4}-\frac{\dot{\nu}\dot{\rho}}{4}\right),\\
R_{rr} &= \left(\frac{\nu''}{2}+\frac{\nu'^2}{4}-\frac{\nu'\rho'}{4}-\frac{\rho'}{r}\right)-\left(\frac{\ddot{\rho}}{2}+\frac{\dot{\rho}^2}{4}-\frac{\dot{\nu}\dot{\rho}}{4}\right)e^{\rho-\nu},\\
R_{0r} &= -\frac{\dot{\rho}}{r},\\
R_{\theta\theta} &= \left[1+\frac{r}{2}\left(\nu'-\rho'\right)\right]e^{-\rho}-1,\\
R_{\phi\phi} &= \sin^2\theta R_{\theta\theta}.
\end{aligned} \tag{12.38}$$

So far we have only discussed the restriction that spherical symmetry places on the solution, and have not sought the actual solution to the Einstein field equation. This we shall do in the following section.

The Einstein equation for the spacetime exterior to the source

Here we wish to find the metric in the region outside a spherically symmetric source. Because the energy–momentum tensor $T_{\mu\nu}$ vanishes in the exterior, the Einstein field equation becomes

$$R_{\mu\nu} = 0. \tag{12.39}$$

Do not be deceived by the superficially simple form of this equation. Keep in mind that the Ricci tensor is a set of a second-order nonlinear differential operators acting on the metric functions, as displayed in (12.38). In this spherical symmetrical case with only two nontrivial scalar functions $\nu(r,t)$ and $\rho(r,t)$, we expect this represents two coupled partial differential equations.

Remark: One should keep in mind that a vanishing Ricci tensor $R_{\mu\nu} = 0$ does not imply a vanishing Riemann tensor $R_{\mu\nu\alpha\beta} = 0$. Namely, an empty space ($T_{\mu\nu} = 0$) does not need to be flat, even though a flat space $R_{\mu\nu\alpha\beta} = 0$ must have a vanishing Ricci tensor. (It may be helpful to compare the situation to the case of a matrix having a vanishing trace. This certainly does not require the entire matrix to vanish.)

Isotropic metric is time independent

Before getting the solution for the two unknown metric functions $g_{00}(r,\, t) \equiv -e^{\nu(r,t)}$ and $g_{rr}(r,t) \equiv e^{\rho(r,t)}$, we first point out that the metric must necessarily be time-independent (the Birkhoff theorem, see Box 12.3)

$$\nu(r,t) = \nu(r) \quad \text{and} \quad \rho(r,t) = \rho(r). \tag{12.40}$$

After substituting in this condition that all t-derivative terms vanish $\dot{\nu} = \dot{\rho} = \ddot{\rho} = 0$, the Einstein vacuum relations in (12.38) yield three component equations:

$R_{00} = 0$:

$$\frac{\nu''}{2} + \frac{\nu'^2}{4} - \frac{\nu'\rho'}{4} + \frac{\nu'}{r} = 0, \tag{12.41}$$

$e^{\rho-\nu}R_{00} + R_{rr} = 0$:

$$\nu' + \rho' = 0, \tag{12.42}$$

$R_{\theta\theta} = 0$:

$$\left[1 + \frac{r}{2}(\nu' - \rho')\right] e^{-\rho} - 1 = 0. \tag{12.43}$$

Actually one of these three equations is redundant. It can be shown that the solution to two equations, for example, (12.42) and (12.43), automatically satisfies the remaining Eq. (12.41).

Box 12.3 The Birkhoff theorem

Theorem: *Every spherically symmetric vacuum solution to $R_{\mu\nu} = 0$ is static. That is, $\dot{\nu} = \dot{\rho} = 0$.*

Proof: That ρ has no time dependence follows simply from the equation $R_{0r} = -\dot{\rho}/r = 0$ in (12.38). That ν has no time dependence can be demonstrated as follows: because ρ and, hence also, ρ' have no t-dependence, the Einstein equation

$$R_{\theta\theta} = \left[1 + \frac{r}{2}(\nu' - \rho')\right] e^{-\rho} - 1 = 0, \tag{12.44}$$

implies that ν' is also time independent (as there is no time dependence in the entire equation). The statement

$$\nu' \equiv \frac{d\nu}{dr} = f(r) \tag{12.45}$$

means that the function ν must depend on the variables r and t, separately:

$$\nu(r,t) = \nu(r) + n(t). \tag{12.46}$$

The appearance of $\nu(r)$ and $n(t)$ in the infinitesimal interval ds^2 has a form so that a possible time-dependence $n(t)$ can be absorbed in a new time variable t':

$$-e^{\nu(r)}\, e^{n(t)}\, c^2dt^2 \equiv -e^{\nu(r)}c^2dt'^2.$$

In terms of these coordinates, the metric functions are time independent. This completes our proof of the Birkhoff theorem. ■

▶ Recall the simple physical argument for the Newtonian analog of the Birkhoff theorem, given at the end of Section 6.1.

▶ Historically, Schwarzschild obtained his solution by explicitly assuming a static spherical source. Only several years later did Birkhoff provide his theorem showing that the solution Schwarzschild obtained was actually valid for an exploding, collapsing, or pulsating spherical star.

Solving the Einstein equation

We now carry out the solution to (12.42) and (12.43). After an integration over r of (12.42), we obtain the equality

$$\nu(r) = -\rho(r), \tag{12.47}$$

where we have set the integration constant to zero by a choice of new time coordinates in exactly the same manner as done in the proof of the Birkhoff theorem (Box 12.3). Because ν and ρ are exponents of the metric scalar functions (12.31), this relation (12.47) translates into

$$-g_{00} = \frac{1}{g_{rr}}. \tag{12.48}$$

(12.42) also allows us to rewrite (12.43) as

$$(1 - r\rho')e^{-\rho} - 1 = 0. \tag{12.49}$$

We can simplify this equation by introducing a new variable:

$$\lambda(r) \equiv e^{-\rho(r)}, \quad \frac{d\lambda}{dr} = -\rho' e^{-\rho}$$

so that (12.49) becomes

$$\frac{d\lambda}{dr} + \frac{\lambda}{r} = \frac{1}{r}, \tag{12.50}$$

which has the general solution $\lambda(r) = \lambda_0(r) + \lambda_1$ where λ_0 is the solution to the homogeneous equation

$$\frac{d\lambda_0}{dr} = -\frac{\lambda_0}{r}. \tag{12.51}$$

This can be solved by straightforward integration, $\ln \lambda_0 = -\ln r + c_0$. It thus implies that the product of $\lambda_0 r$ is a constant, which we label

$$\lambda_0 r \equiv -r^*. \tag{12.52}$$

Combining this with a particular solution of $\lambda_1 = 1$, we have the general solution of

$$\lambda = 1 - \frac{r^*}{r} = \frac{1}{g_{rr}} = -g_{00}, \tag{12.53}$$

where we have used (6.17) and noted that the λ function is just g_{rr}^{-1}. This solution is called the **Schwarzschild metric:**

$$g_{\mu\nu} = \text{diag}\left[\left(-1 + \frac{r^*}{r}\right), \left(1 - \frac{r^*}{r}\right)^{-1}, r^2, r^2 \sin^2\theta\right], \tag{12.54}$$

and is quoted in (6.18). The parameter r^* is then related to Newton's constant and source mass $r^* = 2G_N M/c^2$ through the relation between the metric element and gravitational potential in the Newtonian limit:

$$g_{00} = -\left(1 + \frac{2\Phi}{c^2}\right) = -1 + \frac{2G_N M}{c^2 r} = -1 + \frac{r^*}{r}. \tag{12.55}$$

This Schwarzschild solution (12.54) to the Einstein field equation must be considered as a main achievement of GR in the field of astrophysics. It is an exact solution which corresponds historically to Newton's treatment of the $1/r^2$ force law (12.9) in the classical gravitational theory. Numerous GR applications, from the bending of a light-ray to black holes, are based on this solution (cf. Chapter 6.)

As nonlinear equations are very difficult to solve, it is astonishing that Karl Schwarzschild, the Director of Potsdam Observatory, discovered these exact solutions[3] only two months after Einstein's final formulation of GR at the end of November 1915. At this time Schwarzschild was already in the German army on the Russian front. Tragically by the summer of 1916 he died there (of an illness)—one of the countless victims of the First World War.

12.4 The Einstein equation for cosmology

Cosmological study must be carried out in the framework of GR. The basic dynamical equation is the Einstein equation. In Section 12.4.1 we find the solution of Einstein's equation that is compatible with a 3D space being homogeneous and isotropic as required by the cosmological principle. This solution is the Robertson–Walker metric presented in Chapter 7. The Einstein equation with a Robertson–Walker metric leads to Friedmann equations discussed in Chapter 8. Finally in Section 12.4.3 we show how the Einstein equation can be modified by the addition of the cosmological constant term. The physical implications of such a Λ term have been studied in Chapter 9.

12.4.1 Solution for a homogeneous and isotropic 3D space

The cosmological principle gives us a picture of the universe as a system of "cosmic fluid." It is convenient to pick the coordinate time t to be the proper time of each fluid element. The 4D metric in this comoving coordinate system has the form as discussed in Section 7.3: $g_{\mu\nu} = \text{diag}(-1, g_{ij})$ so that the spacetime separation

$$ds^2 = -c^2dt^2 + dl^2 \tag{12.56}$$

with

$$dl^2 = g_{ij}x^i x^j = [R(t)]^2 d\hat{l}^2, \tag{12.57}$$

where $R(t)$ is the dimensionful scale factor, equal to $a(t)R_0$. The 3D separation $d\hat{l}^2$ is then dimensionless. Previously we argued that the requirement of a homogeneous and isotropic space means that the space must have constant curvature. Then we used the result obtained in Section 4.3.2 for a constant curvature 3D space (heuristically from the result of 2D surfaces of constant curvature):

$$d\hat{l}^2 = \frac{d\xi^2}{1 - k\xi^2} + \xi^2 d\theta^2 + \xi^2 \sin^2\theta d\phi^2 \tag{12.58}$$

with ξ being the dimensionless radial distance. This is the Robertson–Walker metric. Here in this subsection, we shall use the intermediate steps of Section 12.3 (in arriving at the Schwarzschild solution) to provide another derivation of this result (12.58). The purpose is to make it clear that such a metric is indeed the solution of the Einstein equation for a homogeneous and isotropic space.[4]

[3] There is also the Schwarzschild interior solution for the Einstein field equation with $T_{\mu\nu} \neq 0$ (e.g. that for the ideal fluid). Such solutions are relevant for the discussion of gravitational collapse.

[4] It should nevertheless be emphasized the Robertson–Walker metric follows from the symmetry of 3D space, rather from any specific property of gravity as encoded in the Einstein equation. A more rigorous derivation would involve the mathematics of symmetric spaces, Killing vectors, and isometry. See Weinberg, "Gravitation and cosmology," Chapter 13, Wiley, 1972.

Homogeneity and isotropy means that the space must be spherically symmetric with respect to every point in that space. We can work out the metric that satisfies this requirement as follows:

Spherically symmetric with respect to the origin. This means that the metric for the 3D space (ξ, θ, ϕ) should have the form as discussed in Sections 6.1.1 and 12.3.1. Keeping only the spatial part of (12.30), we have

$$d\hat{l}^2 = \hat{g}_{\xi\xi} d\xi^2 + \xi^2 \left(d\theta^2 + \sin^2\theta d\phi^2 \right). \tag{12.59}$$

Birkhoff's theorem (Box 12.3) then implies that the metric element $\hat{g}_{\xi\xi}$ is independent of the coordinate time. (This shows the consistency of our assumption that the reduced metric $\hat{g}_{ij}$, after factoring out the scale factor $a^2(t)$, does not change with time.) We will also follow the previous notation of $\hat{g}_{\xi\xi} \equiv e^{\rho(\xi)}$ as shown in (12.31).

Spherically symmetric with respect to every point. To broaden from the spherical symmetry with respect to one point (the origin) as discussed in Section 12.3.1 to that with respect to every point (as required by homogeneity and isotropy), we demand that the Ricci scalar (for this 3D space), which in general is a function of ξ, be a constant; with some foresight we set it equal to $-6k$,

$$[R^{(3)}] \equiv -6k. \tag{12.60}$$

This just says that the 3D space should be one with constant curvature. We can look up the expression for the Ricci tensor in (12.38), and after setting $\nu = \dot{\nu} = \nu' = \nu'' = \dot{\rho} = \ddot{\rho} = 0$, we obtain the Ricci tensor elements for the 3D space:

$$\begin{aligned} \left[R^{(3)}_{\xi\xi}\right] &= -\frac{1}{\xi}\frac{d\rho}{d\xi}, \qquad \left[R^{(3)}_{\theta\theta}\right] = \left(1 - \frac{\xi}{2}\frac{d\rho}{d\xi}\right) e^{-\rho} - 1, \\ \left[R^{(3)}_{\phi\phi}\right] &= \sin^2\theta \left[R^{(3)}_{\theta\theta}\right], \end{aligned} \tag{12.61}$$

which is to be contracted with the inverse metric $\hat{g}^{ij}$ of (12.59),

$$\hat{g}^{\xi\xi} = e^{-\rho(\xi)}, \quad \hat{g}^{\theta\theta} = \xi^{-2}, \quad \hat{g}^{\phi\phi} = \frac{1}{\xi^2 \sin^2\theta}, \tag{12.62}$$

to obtain the Ricci scalar:

$$\begin{aligned} [R^{(3)}] &= \sum_i [R^{(3)}_{ii}]\hat{g}^{ii} = [R^{(3)}_{\xi\xi}]\hat{g}^{\xi\xi} + 2[R^{(3)}_{\theta\theta}]\hat{g}^{\theta\theta} \\ &= -\frac{e^{-\rho}}{\xi}\frac{d\rho}{d\xi} + \frac{2}{\xi^2}\left[\left(1 - \frac{\xi}{2}\frac{d\rho}{d\xi}\right) e^{-\rho} - 1\right]. \end{aligned}$$

Setting it to $-6k$ as in (12.60)

$$\frac{2}{\xi^2}\frac{d}{d\xi}(\xi e^{-\rho} - \xi) = -6k. \tag{12.63}$$

We can solve this differential equation by straightforward integration

$$\begin{aligned} d(\xi e^{-\rho} - \xi) &= -3k\xi^2 d\xi, \\ (1 - e^{-\rho})\xi &= k\xi^3 + A, \end{aligned} \tag{12.64}$$

where the integration constant $A = 0$, as can be seen in the $\xi = 0$ limit. We obtain the desired solution

$$\hat{g}_{\xi\xi} = e^{\rho(\xi)} = \frac{1}{1 - k\xi^2}. \tag{12.65}$$

Plugging this into (12.59), we have the dimensionless separation in the 3D space as given by (12.58), confirming the heuristic results of (4.46) and (7.43).

12.4.2 Friedmann equations

In this subsection, we shall explicate the exact relation between the Einstein and Friedmann equations used in Chapter 8. In the Einstein equation $G_{\mu\nu} = \kappa T_{\mu\nu}$ (with $\kappa = -8\pi G_{\rm N}/c^4$) for the homogeneous and isotropic universe, the LHS is determined by the Robertson–Walker metric with its two parameters; the curvature constant k and the scale factor $R(t)$. We still need to specify the energy–momentum tensor on the RHS, which must be compatible with cosmological principle. The simplest plausible choice is to take the cosmic fluid as an ideal fluid as discussed in Section 10.3. In special relativity, we have already shown in (12.30) that

$$T_{\mu\nu} = pg_{\mu\nu} + \left(\rho + \frac{p}{c^2}\right) U_\mu U_\nu, \tag{12.66}$$

where p is pressure, ρ mass density, and U^μ 4-velocity field of the fluid. Since there is no derivative, the same form also holds for GR. In the cosmic rest frame (the comoving coordinates) in which each of the fluid element (galaxy) carries its own position label, all the fluid elements are at rest $U^\mu = (c, 0)$. In such a frame with a metric given by $g_{\mu\nu} = \text{diag}(-1, g_{ij})$, the energy–momentum takes on a particularly simple form

$$T_{\mu\nu} = \begin{pmatrix} \rho c^2 & 0 \\ 0 & pg_{ij} \end{pmatrix}. \tag{12.67}$$

The cosmological Friedmann equations are just the Einstein equation with Robertson–Walker metric and with ideal fluid energy–momentum tensor.

1. The $G_{00} = -8\pi G_{\rm N}\rho/c^2$ equation can then be written (again after a long calculation) in terms of the R–W metric elements $R(t)$ and k. We have the first Friedmann equation,

$$\frac{\dot{R}^2(t) + kc^2}{R^2(t)} = \frac{8\pi G_{\rm N}}{3}\rho. \tag{12.68}$$

2. From the $G_{ij} = -8\pi G_{\rm N} p g_{ij}/c^4$ equation, we have the second Friedmann equation,

$$\frac{\ddot{R}(t)}{R(t)} = -\frac{4\pi G_{\rm N}}{c^2}\left(p + \frac{1}{3}\rho c^2\right). \tag{12.69}$$

As we have shown in Chapter 8 these Friedmann equations, because of cosmological principle, have simple Newtonian interpretation. Nevertheless, they must be understood in the context of GR as they still involve the geometric concepts like curvature, etc. The proper view is that they are Einstein equations applied to cosmology.

12.4.3 The Einstein equation with a cosmological constant term

Einstein's desideratum for a static universe led him to modify his original field equation for GR. Given the strong theoretical arguments (cf. Section 12.2) used in arriving at (12.13) and its success in describing gravitation phenomena (at least up to the solar system), how can we go about making such a modification? The possibility is that there is some gravitational feature which is too small to be observed for systems at sub-cosmic scales, but becomes important only on the truly large dimensions. Still, whatever we add to the Einstein equation, it must be compatible with its tensor structure—a symmetric rank-2 tensor that is covariantly constant (i.e. its covariant derivative vanishes). Recall that the Einstein tensor $G_{\mu\nu}$, being a nonlinear second order derivative of the metric, is such a tensor. But the metric tensor $g_{\mu\nu}$ itself is also symmetric, rank-2, and covariantly constant, see (11.32). Thus it is mathematically consistent to include such a term on the LHS of the field equation:

$$G_{\mu\nu} - \Lambda g_{\mu\nu} = \kappa T_{\mu\nu},$$

where

$$\kappa = -\frac{8\pi G_{\mathrm{N}}}{c^4}. \tag{12.70}$$

Λ is some unknown coefficient. The addition will alter the Newtonian limit of the field equation as discussed in Section 12.2.2, and leads to a nonrelativistic equation different from the Newton/Poisson equation, cf. (9.5). This difficulty can, however, be circumvented by assuming that Λ is of such a small size as to be unimportant except for cosmological applications. Hence Λ is called the **cosmological constant.**

While it is more straightforward to see, from a mathematical viewpoint, how the geometry side of Einstein's equation can be modified by this addition, the physical interpretation of this new term can be more readily gleaned if we move it to the energy–momentum side:

$$G_{\mu\nu} = \kappa(T_{\mu\nu} + \kappa^{-1}\Lambda g_{\mu\nu}) = \kappa(T_{\mu\nu} + T^{\Lambda}_{\mu\nu}), \tag{12.71}$$

where $T^{\Lambda}_{\mu\nu} = \kappa^{-1}\Lambda g_{\mu\nu}$ can be called the "vacuum energy tensor." In the absence of ordinary mass/energy distribution $T_{\mu\nu} = 0$ (hence, the vacuum), the source term $T^{\Lambda}_{\mu\nu}$ can still bring about a gravitational field in the form of a nontrivial spacetime curvature.

In the cosmic rest frame (the comoving coordinates) with the velocity field being $U^{\mu} = (c, 0, 0, 0)$ and the metric $g_{\mu\nu} = \mathrm{diag}(-1, g_{ij})$ of (7.37), this vacuum energy–momentum tensor can be written in a form analogous to the conventional ideal fluid stress tensor (12.67):

$$T^{\Lambda}_{\mu\nu} = \frac{\Lambda}{\kappa}\begin{pmatrix} -1 & 0 \\ 0 & g_{ij} \end{pmatrix} \equiv \begin{pmatrix} \rho_{\Lambda}c^2 & 0 \\ 0 & p_{\Lambda}g_{ij} \end{pmatrix}. \tag{12.72}$$

Comparing it to Eq. (12.67), this implies a constant vacuum energy density,

$$\rho_{\Lambda} = -\frac{\Lambda}{\kappa c^2} = \frac{\Lambda c^2}{8\pi G_{\mathrm{N}}}, \tag{12.73}$$

which is the result quoted in Chapter 9 (cf. (9.2)). If we take $\Lambda > 0$, so that $\rho_\Lambda > 0$, it implied a **negative** vacuum pressure:

$$p_\Lambda = -\rho_\Lambda c^2 < 0. \tag{12.74}$$

Thus the cosmological constant corresponds to an energy density which is constant in time and in space. No matter how we change the volume, this energy density is unchanged. As we have discussed in Chapter 9, such negative pressure is the source of gravitational repulsion which can drive the inflationary epoch of the big bang, and can give rise to an universe undergoing an accelerated expansion.

Review questions

1. What is the **principle of general covariance**?
2. Since SR equations are valid only in the absence of gravity, turning SR into GR equations implies the introduction of a gravitational field into relativistic equations. If the physics equation is known in the special relativistic limit, how does one turn such an SR equation into a general relativistic one? Also discuss the difference of coordinate symmetries involved in SR and GR.
3. How can one "deduce" the GR equation of motion from that of SR?
4. Why did Einstein expect the relativistic gravitational field equation to have the form of $[\hat{O}g] = \kappa[T]$ with the LHS being a covariantly constant symmetric tensor of rank-2 involving $(\partial^2 g)$ as well as $(\partial g)^2$ terms?
5. Write out the two equivalent versions of the Einstein field equation, with the coupling expressed in terms of Newton's constant.
6. How can we use the result of a metric for a spherical symmetric space to derive that for a space that is homogeneous and isotropic?
7. What is the relation between the Friedmann equations and the Einstein equation?
8. What are the mathematical properties of the cosmological constant term that allow it to be added to the Einstein equation?
9. Write out the Einstein equation with the Λ term. Explain why such a term can be interpreted as the vacuum energy–momentum source of gravity.

Problems

(12.1) **Another derivation of geodesic equation** Starting from the no-force condition $(d\mathbf{p})^\mu = Dp^\mu = 0$ in a curved spacetime with the 4-momentum $p^\mu = mU^\mu$, show how to arrive at the geodesic equation by using the covariant derivative expression of (11.45).

(12.2) **Vacuum Einstein equations**

(a) Show that a vanishing Einstein tensor implies a vanishing Ricci tensor.

(b) The three Eqs (12.41) to (12.43) are redundant. Show that the solution to two equations, for example, Eqs (12.42) and (12.43), automatically satisfies the remaining Eq. (12.41).

(12.3) **Friedmann equations and energy conservation** Show that energy conservation statement (8.3) that results from the linear combination of these two Friedmann Eqs (12.68) and (12.69) can also be derived directly from the energy–momentum conservation equation of $D_\mu T^{\mu\nu} = 0$.

(12.4) **The equation of geodesic deviation** We have derived an expression for the curvature (11.58) on pure geometric considerations. A more physics approach would be to seek the GR generalization of tidal forces as discussed in Section 5.3. Following exactly the same steps used to derive the Newtonian deviation equation in Box 5.2, one can obtain its GR version, called the equation of geodesic deviation,

$$\frac{D^2 s^\mu}{D\tau^2} = -R^{\mu}_{\ \alpha\nu\beta} s^\nu \frac{dx^\alpha}{d\tau}\frac{dx^\beta}{d\tau}. \tag{12.75}$$

Namely, the tensor of gravitational potential's second derivative is replaced by the Riemann curvature tensor (11.58). This derivation requires a careful discussion of the second derivative along a geodesic curve, cf. (11.46).

(12.5) **From geodesic deviation to NR tidal forces** Show that the equation of geodesic deviation (12.75) reduces to the Newtonian deviation Eq. (5.32) in the Newtonian limit. The GR Eq. (12.75) is reduced to

$$\frac{d^2 s^i}{dt^2} = -c^2 R^j_{\ 0j0} s^j$$

in the NR limit of a slow moving particle with 4-velocity of $dx^\alpha/d\tau \simeq (c, 0, 0, 0)$. We have also set $s^0 = 0$ because we are comparing the two particles' acceleration at the same time. Thus, (5.32) can be recovered by showing the relation

$$R^j_{\ 0j0} = \frac{1}{c^2} \frac{\partial^2 \Phi}{\partial x^i \partial x^j}$$

in the Newtonian limit. You are asked to prove this limit expression for the Riemann curvature.

13 Linearized theory and gravitational waves

- In the weak-field limit Einstein's equation can be linearized and takes on the familiar wave equation form.
- Gravitational waves may be viewed as ripples of curvature propagating in a background of flat spacetime.
- The strategy of detecting such tidal forces by a gravitational wave interferometer is outlined.
- The rate of energy loss due to quadrupole radiation by a circulating binary system is calculated, and found in excellent agreement with the observed orbit decay rate of the relativistic binary pulsar PSR 1913+16.

Newton's theory of gravitation is a static theory. The Newtonian field due to a source is established instantaneously. Thus, while the field has nontrivial dependence on the spatial coordinates, it does not depend on time. Einstein's theory, being relativistic, is symmetric with respect to space and time. Just like Maxwell's theory, it has the feature that a field propagates outward from the source with a finite speed. In this chapter, we study the case of a weak gravitational field. This approximation linearizes the Einstein theory. In this limit, a gravitational wave may be viewed as small curvature ripples propagating in a background of flat spacetime. It is a transverse wave having two independent polarization states, traveling at the speed of light.

Because gravitational interaction is so weak, any significant emission of gravitational radiation can come only from a strong field region involving dynamics that directly reflects GR physics. Once gravitational waves are emitted, they will not scatter and they propagate out undisturbed from the inner core of an imploding star, from the arena of black hole formation, and from the earliest moments of the universe, etc. They come from regions which are usually obscured in electromagnetic, even neutrino astronomy: gravitational waves can provide us with a new window into astrophysical phenomena.

These ripples of curvature can be detected as tidal forces. We provide an outline of the detection strategy using the gravitational wave interferometers, which can measure the minute compression and elongation of orthogonal lengths that are caused by the passage of such a wave. In the final section, we present the indirect, but convincing, evidence for the existence of gravitational waves as predicted by general relativity (GR). This came from the observation, spanning more than 25 years, of orbital motion of the relativistic binary pulsar[1]system: PSR 1913+16. Even though the binary pair is 5 kpc away, the basic parameters of the system can be deduced by carefully monitoring the radio pulses emitted by the pulsar, which effectively acted as an accurate and stable clock. From

[1] A pulsar is a magnetized neutron star whose rapid rotation generates a circulating plasma that serves as a source of beamed radio waves detectable on earth as periodic pulses.

this record we can verify a number of GR effects. In particular the orbit period is observed to decrease. According to GR, this is brought about by the gravitational wave quadrupole radiation from the system. The observed rate is in splendid agreement with the prediction by Einstein's theory.

13.1 The linearized Einstein theory

The production of gravitational waves usually involves strong-field situations, but, because of the weakness of the gravitational interaction, the produced gravitational waves are only tiny displacements of the flat spacetime metric. Thus, it is entirely adequate for the description of gravity wave to restrict ourselves to the situation of a weak gravitation field. The Newtonian limit corresponds to nonrelativistic motion in a weak static field. Here we remove the restriction of slow motion and allow for a time-dependent field. In a weak field, the metric is almost Minkowskian:

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu} \equiv g^{(1)}_{\mu\nu}, \tag{13.1}$$

where $|h_{\mu\nu}| \ll 1$ everywhere in spacetime. Thus, we will keep only first-order terms in $h_{\mu\nu}$, and denote the relevant quantities with a superscript $^{(1)}$. The idea is that slightly curved coordinate systems exist and they are suitable coordinates to use in the weak-field situation. We can still make coordinate transformations among such systems—from one slightly curved one to another. In particular we can make a "background Lorentz transformation." Distinguishing the indices, $\{\mu\}$ vs. $\{\mu'\}$, to indicate the untransformed and transformed coordinates, we have

$$x^{\mu} \rightarrow x^{\mu'} = [\mathbf{L}]^{\mu'}_{\nu} x^{\nu}, \tag{13.2}$$

where $[\mathbf{L}]$ is the position-independent Lorentz transformation of special relativity (cf. (10.9) and (10.10)). The key property of such transformations is that they keep the Minkowski metric invariant, see (10.13),

$$[\bar{\mathbf{L}}]^{\mu}_{\alpha'}[\bar{\mathbf{L}}]^{\nu}_{\beta'}\eta_{\mu\nu} = \eta_{\alpha'\beta'}. \tag{13.3}$$

This leads to the transformation of the full metric as

$$[\bar{\mathbf{L}}]^{\mu}_{\alpha'}[\bar{\mathbf{L}}]^{\nu}_{\beta'}g^{(1)}_{\mu\nu} = \eta_{\alpha'\beta'} + [\bar{\mathbf{L}}]^{\mu}_{\alpha'}[\bar{\mathbf{L}}]^{\nu}_{\beta'}h_{\mu\nu} = g^{(1)}_{\alpha'\beta'}. \tag{13.4}$$

Thus

$$h_{\alpha'\beta'} = [\bar{\mathbf{L}}]^{\mu}_{\alpha'}[\bar{\mathbf{L}}]^{\nu}_{\beta'}h_{\mu\nu}. \tag{13.5}$$

Namely, $h_{\mu\nu}$ is just a Lorentz tensor. Thus, this part of the metric can be taken as a tensor defined on a flat Minkowski spacetime. Since the nontrivial physics is contained in $h_{\mu\nu}$, we can have the convenient picture of a weak gravitational field as being described by this symmetric field $h_{\mu\nu}$ in a flat spacetime.[2]

Dropping higher order terms of $h_{\mu\nu}$, we have the Riemann curvature tensor

$$R^{(1)}_{\alpha\mu\beta\nu} = \frac{1}{2}(\partial_\alpha\partial_\nu h_{\mu\beta} + \partial_\mu\partial_\beta h_{\alpha\nu} - \partial_\alpha\partial_\beta h_{\mu\nu} - \partial_\mu\partial_\nu h_{\alpha\beta}), \tag{13.6}$$

the Ricci tensor

$$\begin{aligned} R^{(1)}_{\mu\nu} &= \eta^{\alpha\beta}R^{(1)}_{\alpha\mu\beta\nu} \\ &= \frac{1}{2}(\partial_\alpha\partial_\nu h^{\alpha}_{\mu} + \partial_\mu\partial_\alpha h^{\alpha}_{\nu} - \Box h_{\mu\nu} - \partial_\mu\partial_\nu h), \end{aligned} \tag{13.7}$$

[2] Eventually in a quantum description, $h_{\mu\nu}$ is a field for the spin-2 gravitons, and the perturbative description of gravitational interactions as due to the exchanges of massless gravitons.

and the Ricci scalar

$$R^{(1)} = \partial_\mu \partial_\nu h^{\mu\nu} - \Box h, \tag{13.8}$$

where $\Box = \partial_\mu \partial^\mu$ and $h = h^\mu_\mu$ is the trace. Clearly the resultant Einstein tensor

$$G^{(1)}_{\mu\nu} = R^{(1)}_{\mu\nu} - \frac{1}{2} R^{(1)} \eta_{\mu\nu} \tag{13.9}$$

is also linear in $h_{\mu\nu}$, and so is the Einstein equation:

$$G^{(1)}_{\mu\nu} = -\frac{8\pi G_{\mathrm{N}}}{c^4} T^{(0)}_{\mu\nu}. \tag{13.10}$$

NB: On the right-hand side (RHS) the energy-momentum tensor $T^{(0)}_{\mu\nu}$ has no $h_{\mu\nu}$ dependence because $T_{\mu\nu}$ must already be small $T_{\mu\nu} = O(h_{\mu\nu})$ to be consistent with a spacetime being slightly curved, and its conservation is expressed as

$$\partial^\mu T^{(0)}_{\mu\nu} = 0 \tag{13.11}$$

in terms of ordinary derivatives.

13.1.1 The coordinate change called gauge transformation

In the following, we shall make coordinate transformations so that the linearized Einstein Eq. (13.10) can be written more compactly in terms of $h_{\mu\nu}$. This class of coordinate transformations (within the slightly curved spacetime) is called, collectively, **gauge transformations** because of their close resemblance to the electromagnetic gauge transformations. Consider a small shift of the position vector:

$$x^{\mu'} = x^\mu + \chi^\mu(x), \tag{13.12}$$

where $\chi^\mu(x)$ are four arbitrary small functions. Collectively they are called the "vector gauge function" (as opposed to the scalar gauge function in electromagnetic gauge transformation, cf. Box 10.3). Clearly this is not a tensor equation, as indices do not match on the two sides. (Our notation indicates the relation of the position vector as labeled by the transformed and pre-transformed coordinates.) The smallness of the shift $\chi \ll x$ means

$$|\partial_\mu \chi^\nu| \ll 1. \tag{13.13}$$

The corresponding transformation (for the contravariant components) can be obtained by differentiating (13.12):

$$\frac{\partial x^{\mu'}}{\partial x^\alpha} = \delta^\mu_\alpha + \partial_\alpha \chi^\mu. \tag{13.14}$$

This also implies an inverse transformation of

$$\frac{\partial x^\mu}{\partial x^{\alpha'}} = \delta^\mu_\alpha - \partial_\alpha \chi^\mu + O(|\partial \chi|^2). \tag{13.15}$$

Apply it to the metric tensor:

$$\begin{aligned} g^{(1)}_{\alpha'\beta'} &= \frac{\partial x^\mu}{\partial x^{\alpha'}} \frac{\partial x^\nu}{\partial x^{\beta'}} g^{(1)}_{\mu\nu} \\ &= \delta^\mu_\alpha \delta^\nu_\beta g^{(1)}_{\mu\nu} - \partial_\alpha \chi^\mu \eta_{\mu\beta} - \partial_\beta \chi^\nu \eta_{\nu\alpha} \\ &= g^{(1)}_{\alpha\beta} - \partial_\alpha \chi_\beta - \partial_\beta \chi_\alpha, \end{aligned} \tag{13.16}$$

Table 13.1 Analog between the electromagnetic and linearized gravitational field theory

	Electromagnetism	Linearized gravity
Source	j^{μ}	$T^{\mu\nu}$
Conservation law	$\partial_{\mu} j^{\mu} = 0$	$\partial_{\mu} T^{\mu\nu} = 0$
Field	A_{μ}	$h_{\mu\nu}$
Gauge transformation	$A_{\mu} \to A_{\mu} - \partial_{\mu}\chi$	$h_{\mu\nu} \to h_{\mu\nu} - \partial_{\mu}\chi_{\nu} - \partial_{\nu}\chi_{\mu}$
Preferred gauge	$\partial^{\mu} A_{\mu} = 0$	$\partial^{\mu}\bar{h}_{\mu\nu} = 0$
(Lorentz gauge)		$\bar{h}_{\mu\nu} = h_{\mu\nu} - \frac{1}{2} h \eta_{\mu\nu}$
Field equation in the preferred gauge	$\Box A_{\mu} = \frac{4\pi}{c} j_{\mu}$	$\Box \bar{h}_{\mu\nu} = (16\pi G_{\mathrm{N}}/c^4) T_{\mu\nu}$

where $\chi_{\alpha} = \chi^{\mu}\eta_{\mu\alpha}$. Expressing both sides in terms of $h_{\alpha\beta}$, we have the gauge transformation of the perturbation field

$$h_{\alpha'\beta'} = h_{\alpha\beta} - \partial_{\alpha}\chi_{\beta} - \partial_{\beta}\chi_{\alpha}, \tag{13.17}$$

which closely resembles the transformation (10.69) for the electromagnetic 4-vector potential $A_{\alpha}(x)$ (Table 13.1).

13.1.2 The wave equation in the Lorentz gauge

Just as in electromagnetism, one can streamline some calculation by an appropriate choice of gauge conditions. Here this means that a particular choice of coordinates can simplify the field equation formalism for gravitational waves. We are interested in the coordinate system (cf. Problem 13.1) for which the **Lorentz gauge** (also known as the **harmonic gauge**) condition holds:

$$\partial^{\mu}\bar{h}_{\mu\nu} = 0, \tag{13.18}$$

where

$$\bar{h}_{\mu\nu} = h_{\mu\nu} - \frac{h}{2}\eta_{\mu\nu} \tag{13.19}$$

has a trace of opposite sign, $\bar{h}^{\mu}_{\mu} \equiv \bar{h} = -h$. It can then be shown that this **trace reversed perturbation** has the gauge transformation of

$$\bar{h}_{\alpha'\beta'} = \bar{h}_{\alpha\beta} - \partial_{\alpha}\chi_{\beta} - \partial_{\beta}\chi_{\alpha} + \eta_{\alpha\beta}(\partial^{\mu}\chi_{\mu}). \tag{13.20}$$

From (13.18) and (13.19), we have the Lorentz gauge relation $\partial^{\mu}h_{\mu\nu} = \frac{1}{2}\partial_{\nu}h$, which implies, in (13.7) and (13.8), a simplified Ricci tensor $R^{(1)}_{\mu\nu} = -\frac{1}{2}\Box h_{\mu\nu}$, scalar $R^{(1)} = -\frac{1}{2}\Box h$ thus turning the linearized Einstein Eq. (13.10) into the form of a standard wave equation:

$$\Box\bar{h}_{\mu\nu} = \frac{16\pi G_{\mathrm{N}}}{c^4} T^{(0)}_{\mu\nu}. \tag{13.21}$$

Its retarded field solution

$$\bar{h}_{\mu\nu}(\mathbf{x}, t) = \frac{4G_{\mathrm{N}}}{c^4}\int d^3\mathbf{x}' \frac{T^{(0)}_{\mu\nu}(\mathbf{x}', t - |\mathbf{x} - \mathbf{x}'|/c)}{|\mathbf{x} - \mathbf{x}'|} \tag{13.22}$$

is certainly compatible with the gauge condition $\partial^{\mu}\bar{h}_{\mu\nu} = 0$ because of energy-momentum conservation (13.11).

To reiterate, in this linear approximation of the Einstein theory, the metric perturbation $h_{\mu\nu}$ may be regarded as the symmetric field of gravity waves propagating in the background of a flat spacetime. A comparison of the linearized Einstein theory with the familiar electromagnetic equations can be instructive. Such an analog is presented in Table 13.1.

13.2 Plane waves and the polarization tensor

We shall first consider the propagation of a gravitational wave in vacuum. Such ripples in the metric can always be regarded as a superposition of plane waves. A gravity wave has two independent polarization states. Their explicit form will be displayed in a particular coordinate system, the transverse-traceless (T T) gauge.

Plane waves

The linearized Einstein equation in vacuum, (13.21) with $T_{\mu\nu}^{(0)} = 0$, is

$$\Box \bar{h}_{\mu\nu} = 0. \tag{13.23}$$

Because the trace $\bar{h} = -h$ satisfies the same wave equation, we also have

$$\Box h_{\mu\nu} = 0. \tag{13.24}$$

Consider the plane wave solution in the form of

$$h_{\mu\nu}(x) = \epsilon_{\mu\nu} e^{ik_\alpha x^\alpha}, \tag{13.25}$$

where $\epsilon_{\mu\nu}$, the polarization tensor of the gravitational wave, is a set of constants forming a symmetric tensor

$$\epsilon_{\mu\nu} = \epsilon_{\nu\mu} \tag{13.26}$$

and k^α is the 4-wavevector $k^\alpha = (\omega/c, \vec{k})$. Substituting (13.25) into (13.24), we obtain $k^2 \epsilon_{\mu\nu} e^{ikx} = 0$; thus the wavevector must be a null-vector

$$k^2 = k_\alpha k^\alpha = -\frac{\omega^2}{c^2} + \vec{k}^2 = 0. \tag{13.27}$$

Gravitational waves propagate at the same speed $\omega/|\vec{k}| = c$ as electromagnetic waves. Furthermore, because the wave Eq. (13.24) is valid only in the coordinates satisfying the Lorentz gauge condition (13.18), the polarization tensor must be "transverse":

$$k^\mu \epsilon_{\mu\nu} = 0. \tag{13.28}$$

The transverse-traceless gauge

There is still some residual gauge freedom left: one can make further coordinate gauge transformations as long as the transverse condition (13.28) is not violated. This requires that the associated gauge vector function χ_μ be constrained by the condition:

$$\Box \chi_\mu = 0. \tag{13.29}$$

Such coordinate freedom can be used to simplify the polarization tensor (see Problem 13.1): one can pick $\epsilon_{\mu\nu}$ to be traceless,

$$\epsilon^{\mu}_{\mu} = 0, \tag{13.30}$$

as well as

$$\epsilon_{\mu 0} = \epsilon_{0\mu} = 0. \tag{13.31}$$

This particular choice of coordinates is called the "transverse-traceless gauge," which is a subset of coordinates satisfying the Lorentz gauge condition.

The 4×4 symmetric polarization matrix $\epsilon_{\mu\nu}$ has 10 independent elements. Equations (13.28), (13.30), and (13.31) which superficially represent 9 conditions actually fix only 8 parameters because the condition $k^{\mu}\epsilon_{\mu 0} = 0$ is trivially satisfied by (13.31). Thus $\epsilon_{\mu\nu}$ has only two independent elements. The gravitational wave has two independent polarization states. Let us display them. Consider a wave propagating in the z direction $k^{\alpha} = (\omega, 0, 0, \omega)/c$, the transversality condition together with (13.31) implies that $\omega\epsilon_{3\nu} = 0$, or $\epsilon_{3\nu} = \epsilon_{\nu 3} = 0$. Together with the conditions (13.30) and (13.31), the metric perturbation has the form

$$h_{\mu\nu}(z,t) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & h_{+} & h_{\times} & 0 \\ 0 & h_{\times} & -h_{+} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} e^{i\omega(z-ct)/c}. \tag{13.32}$$

The two polarization states can be taken to be

$$\epsilon^{\mu\nu}_{(+)} = h_{+} \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \epsilon^{\mu\nu}_{(\times)} = h_{\times} \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

with h_{+} and $h_{\times}$ being the respective "plus" and "cross" amplitudes.

13.3 Gravitational wave detection

The coordinate-independent feature of any gravitational field is its tidal effect. Thus, the detection of gravitational waves involves the recording of minute changes in the relative positions of a set of test particles. In this section, we shall first deduce the oscillatory pattern of such displacements, then briefly describe the principle underlying the gravitational wave interferometer as detector of such ripples in spacetime.

13.3.1 Effect of gravitational waves on test particles

Consider a free particle before its encounter with a gravitational wave. It is at rest with a 4-velocity $U^{\mu} = (c, 0, 0, 0)$. The effect of the gravitational wave on this test particle is determined by the geodesic equation

$$\frac{dU^{\mu}}{d\tau}+\Gamma^{\mu}_{\nu\lambda}U^{\nu}U^{\lambda}=0. \tag{13.33}$$

Since only U^0 is nonvanishing at the beginning, it reduces to an expression for the initial acceleration of

$$\left(\frac{dU^{\mu}}{d\tau}\right)_0=-c^2\Gamma^{\mu}_{00}. \tag{13.34}$$

The Christoffel symbols on the RHS

$$\Gamma^{\mu}_{00}=\frac{1}{2}\eta^{\mu\nu}(\partial_0 h_{\nu 0}+\partial_0 h_{0\nu}-\partial_\nu h_{00}), \tag{13.35}$$

actually vanish because the metric perturbation $h_{\mu\nu}$ has, in the TT gauge, polarization components of $\epsilon_{\nu 0}=\epsilon_{0\nu}=\epsilon_{00}=0$. The vanishing of initial acceleration means that the particle will be at rest a moment later. Repeating the same argument for later moments, we find the particle at rest for all times. In this way we conclude

$$\frac{dU^{\mu}}{d\tau}=0. \tag{13.36}$$

The particle is stationary with respect to the chosen coordinate system—the TT gauge coordinate labels stay attached to the particle. Thus one cannot discover any gravitational field effect on a single particle. This is compatible with our expectation, based on the equivalence principle (EP), that gravity can always be transformed away at a point by an appropriate choice of coordinates. We need to examine the relative motion of at least two particles in order to detect the oncoming change in the curvature of spacetime.

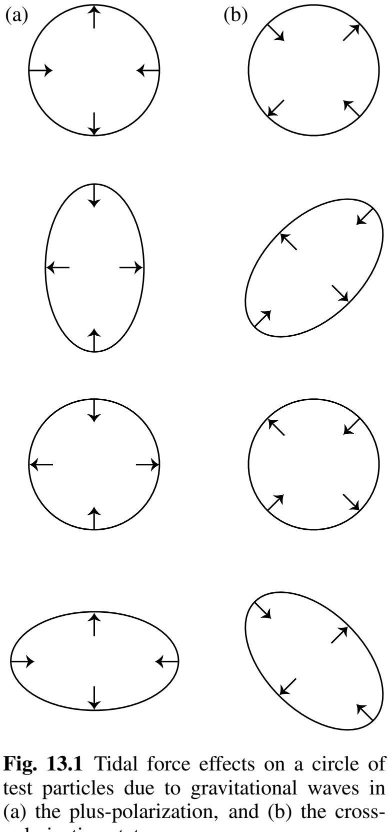

Fig. 13.1 Tidal force effects on a circle of test particles due to gravitational waves in (a) the plus-polarization, and (b) the cross-polarization states.

Consider the effect of a gravitational wave with "plus-polarization" $\epsilon^{\mu\nu}_{(+)}$ on two test particles at rest: one at the origin and the other located at an infinitesimal distance ξ away on the x-axis, hence at an infinitesimally small separation $dx^{\mu}=(0,\xi,0,0)$. This translates into a proper separation of

$$\begin{aligned}ds=\sqrt{g_{\mu\nu}dx^{\mu}dx^{\nu}}=\sqrt{g_{11}}\xi\simeq\left[\eta_{11}+\frac{1}{2}h_{11}\right]\xi\\ =\left[1+\frac{1}{2}h_{+}\sin\omega(t-z/c)\right]\xi\end{aligned} \tag{13.37}$$

showing that the proper distance does change with time. Similarly for two particles separated along the y-axis $dx^{\mu}=(0,0,\xi,0)$ the effect of the gravitational wave is to alter the separation according to

$$ds=\left[1-\frac{1}{2}h_{+}\sin\omega(t-z/c)\right]\xi. \tag{13.38}$$

Thus, the separation along the x direction is elongated while along the y direction compressed. There is no change in the longitudinal separation along the z direction. Just like the electromagnetic waves, gravitational radiation is a transverse field. To better exhibit this pattern of relative displacement we illustrate in Fig. 13.1(a) the effect of a plus-polarized wave, but instead of impinging on two particles as discussed above, acting on a set of test particles when the second particle is replaced by a circle of particles with the first test particle at the center. The outcome that generalizes (13.37) and (13.38) is shown through the wave's one cycle of oscillation.

The effect of a wave with cross-polarization $\epsilon^{\mu\nu}_{(\times)}$ on two particles with a differential interval of $dx^{\mu} = (0, 1, \pm 1, 0)\,\xi/\sqrt{2}$ alters the proper separation as $ds = [1 \pm \frac{1}{2}h_{\times} \sin\omega(t - z/c)]\xi$. The generalization to a circle of particles through one cycle of oscillation is shown in Fig. 13.1(b), which is just a 45° rotation of the plus-polarized wave result of Fig. 13.1(a). While the two independent polarization directions of an electromagnetic wave are at 90° from each other, those of a gravity wave are at 45°. This is related to the feature that, in the dual description of wave as streaming particles, the associated particles of these waves have different intrinsic angular momenta: the photon has spin 1 while the graviton has spin 2. It is also instructive to compare the tidal force effects on such test-particles' relative displacement in response to an oncoming oscillatory gravitational field to that of a static gravitational field as discussed in Section 5.3.1.

13.3.2 Gravitational wave interferometers

A gravitational wave can be thought of as a propagating metric, affecting distance measurements. Thus, as a wave passes through, the distance between two test masses changes with time. The fractional change $l^{-1}\delta l$, called **strain**, is directly related to the wave amplitude h. Although we can obtain an expression of h by using the two energy-flux results (13.55) and (13.66), at this stage it is more instructive to get an estimate by the "hand-waving" argument given in the following paragraph.

The separation between two test masses are given by the equation of geodesic deviation (cf. Problems 12.4 and 12.5), but we shall estimate it by using the simpler Newtonian deviation equation of (5.32), which expresses the acceleration per unit separation by the second derivative of the gravitational potential. We assume that the relativistic effect can be included by a multiplicative factor. The Newtonian potential for a spherical source is $\Phi = -G_{\mathrm{N}}Mr^{-1}$. A gravitational wave propagating in the z direction is a disturbance in the gravitational field:

$$\delta\Phi = -\epsilon\frac{G_{\mathrm{N}}M}{r}\sin(kz - \omega t), \tag{13.39}$$

where $k = \omega/c$. A dimensionless factor of ϵ has been inserted to represent the relativistic correction. The second derivative can be approximated by

$$\frac{\partial^2}{\partial z^2}\delta\Phi = \epsilon\frac{G_{\mathrm{N}}M}{rc^2}\omega^2\sin(kz - \omega t), \tag{13.40}$$

where we have dropped subleading terms coming from differentiation of the r^{-1} factor. This being the acceleration as given in (5.32), the separation amplitude (for the time interval ω^{-1}) per unit separation is then given by

$$h = \left(\frac{\delta s}{s}\right)_{\mathrm{amp}} = \epsilon\frac{G_{\mathrm{N}}M}{rc^2}. \tag{13.41}$$

A similar approximation of the radiation formula (13.66) suggests that the relativistic correction factor ϵ as being the nonspherical velocity squared $(v/c)^2$. The first generation of gravitational wave interferometers have been set up with the aim of detecting gravitation wave emission by neutron stars from the richest source of galaxies in our neighboring part of the universe, the Virgo cluster at $r \approx 15$ Mpc distance away. Thus, even for a sizable $\epsilon = O(10^{-1})$ from a solar mass source $M = M_{\odot}$ the expected strain is only $h = O(10^{-21})$. For two test masses separated by a distance of 10 km the gravity wave induced separation is still one hundredth of a nuclear size dimension. This has been described as showing that spacetime is a very stiff medium, as a large amount of energy can still bring about a tiny disturbance in the spacetime metric. This fact poses great challenge to experimental observation of gravitational waves.

The above discussion makes it clear that one needs to design sensitive detectors to measure the minute length changes between test masses over long distances. Several detectors have been constructed based on the Michelson interferometer configuration (Fig. 13.2). The test masses are mirrors suspended to isolate them from external perturbation forces. Light from a laser source is divided into the two arms by a beam splitter. The light entering into an arm of length L is reflected back and forth in a Fabry–Perot cavity for n times so that the optical length is greatly increased and the storage time is $n(L/c) = \Delta t_n$. The return light-beams from the two arms are combined after they pass through the beam splitter again. By choosing the path length properly, the optical electric field can be made to vanish (destructive interference) at the photodetector. Once adjusted this way, a stretch in one arm and a compression in the other, when induced by the polarization of a passing gravitational wave, will change the optical field at the photodetector in proportion to the product of the field times the wave amplitude. Such an interferometer should be uniformly sensitive to wave frequencies less than $\frac{1}{4}\Delta t_n^{-1}$ (and a loss of sensitivity to higher frequencies). The basic principle to achieve high sensitivity is based on the idea that most of the perturbation noise forces are independent of the baseline lengths while the gravitational-wave displacement grows with the baseline.

The Laser Interferometer Gravitational Observatory (LIGO) is comprised of two sites: one at the Hanford Reservation in Central Washington (Fig. 13.3) housing two interferometers one 2 km- and another 4 km-long arms, while

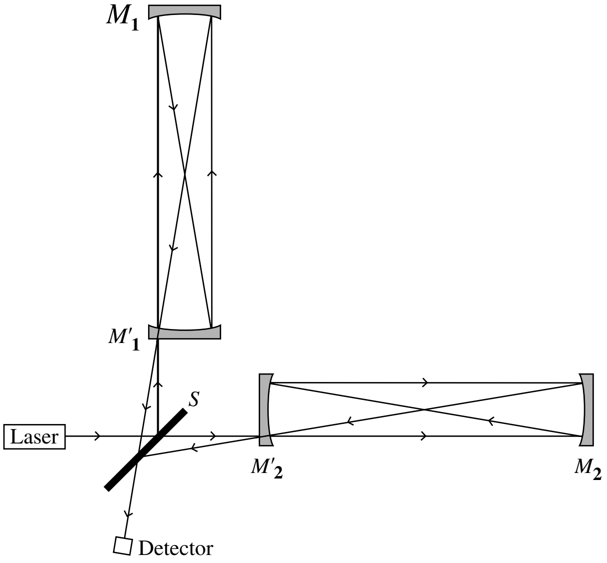

Fig. 13.2 Schematic diagram for gravitational wave Michelson interferometer. The four mirrors $M_{1,2}$, $M'_{1,2}$ and the beam splitter mirror are freely suspended. The two arms are optical cavities that increase the optical paths by many factors. A minute length change of the two arms, one expands and the other contracts, will show up as changes in fringe pattern of the detected light.

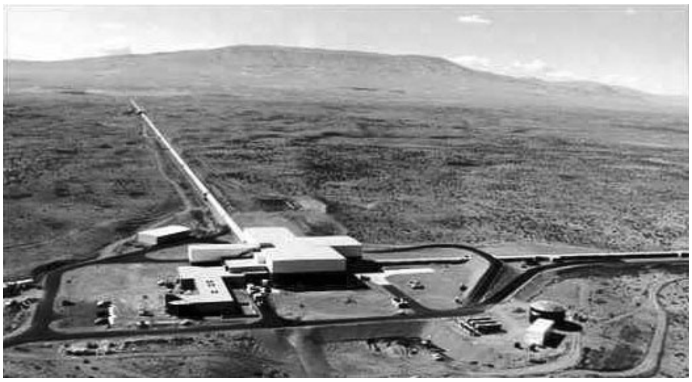

Fig. 13.3 LIGO Hanford Observatory in Washington state.

the other site at Livingston Parish, Louisiana. The three interferometers are being operated in coincidence so that the signal can be confirmed by data from all three sites. Other gravitational wave interferometers in operation are the French/Italian VIRGO project, the German/Scottish GEO project, and the Japanese TAMA project. Furthermore, study is underway both at the European Space Agency and NASA for the launching of three spacecraft placed in solar orbit with one AU radius, trailing the earth by 20°. The spacecraft are located at the corners of an equilateral triangle with sides 5×10^6 km long. Laser Interferometer Space Antenna (LISA) consists of single-pass interferometers, set up to observe a gravitational wave at low frequencies (from 10^{-5} to 1 Hz). This spectrum range is expected to include signals from several interesting interactions of black holes at cosmological distances.

Besides planning and building ever larger scale gravitational wave detectors, a major effort by the theoretical community in relativity is involved in the difficult task of calculating wave shapes in various strong gravity situations (e.g. neutron-star/neutron-star collision, black hole mergers, etc.) to guide the detection and comparison of theory with experimental observations.

13.4 Evidence for gravitational wave

Although, as of this writing, there has not been any generally accepted proof for a direct detection of a gravitational wave, there is nevertheless convincing, albeit indirect, evidence for the existence of such wave as predicted by Einstein's theory. Just as any shaking of electric charges produces electromagnetic waves, a shaking of masses will result in the generation of a gravitational wave, which carries away energy. A system of orbiting binary stars thus loses energy and this results in a decrease of its orbit period. In the following two subsections we present the formula that relates the energy flux of a gravitational wave to the metric perturbation field $h_{\mu\nu}$; and then calculate $h_{\mu\nu}$ produced by the quadrupole radiation as given by the linearized Einstein equation. In the final subsection we present the relativistic binary pulsar system showing that its decrease of orbit period due to gravitational wave radiation is in excellent agreement with what is predicted by GR.

13.4.1 Energy flux in linearized gravitational waves

In the linearized Einstein theory, gravitational waves are regarded as small curvature ripples propagating in a background of flat spacetime. But gravity waves, just like electromagnetic waves, carry energy and momentum; they will in turn produce additional curvature in the background spacetime. Thus we should have a slightly curved background and (13.1) should be generalized to

$$g_{\mu\nu} = g^{(b)}_{\mu\nu} + h_{\mu\nu}, \tag{13.42}$$

where $g^{(b)}_{\mu\nu} = \eta_{\mu\nu} + O(h^2)$ is the background metric. The Ricci tensor can similarly be decomposed as

$$R_{\mu\nu} = R^{(b)}_{\mu\nu} + R^{(1)}_{\mu\nu} + R^{(2)}_{\mu\nu} + \cdots,$$

where $R^{(n)}_{\mu\nu} = O(h^n)$ with $n = 1, 2, \ldots$. In the free space the Einstein equation being $R_{\mu\nu} = 0$, terms corresponding to different orders of metric perturbation on the RHS must vanish separately:

$$R^{(1)}_{\mu\nu} = 0, \tag{13.43}$$

which is just (13.10) for the free space with $T_{\mu\nu} = 0$, and

$$R^{(b)}_{\mu\nu} + R^{(2)}_{\mu\nu} = 0 \tag{13.44}$$

because both terms are quadratic in the metric perturbation $O(h^2)$. The energy momentum tensor carried by the gravity wave $t_{\mu\nu}$ provides the slight curvature of the background spacetime. It must therefore be related to the background Ricci tensor by way of the Einstein Eq. (12.14) at this order,

$$R^{(b)}_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}R^{(b)} = -\frac{8\pi G_{\mathrm{N}}}{c^4}t_{\mu\nu}.$$

Thus $t_{\mu\nu}$ is fixed by $R^{(b)}_{\mu\nu}$, which in turn is related to $R^{(2)}_{\mu\nu}$ by way of (13.44). This allows us to calculate $t_{\mu\nu}$ through the second-order Ricci tensor and scalar:

$$t_{\mu\nu} = \frac{c^4}{8\pi G_{\mathrm{N}}}\left(R^{(2)}_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}R^{(2)}\right). \tag{13.45}$$

Before carrying out the calculation of $t_{\mu\nu}$, we should clarify one point: the concept of **local** energy of a gravitational field does not exist. Namely, one cannot specify the gravitational energy at any single point in space. This is so because the energy being a coordinate-independent function of field, one can always, according to the EP, find a coordinate (the local inertial frame) where the gravity field vanishes locally. Saying it in another way, just as in electromagnetism, we expect the energy density to be proportional to the square of the potential's first derivative. But, according to the flatness theorem, the first derivative of the metric vanishes in the local inertial frame. Thus we cannot speak of gravity's local energy. Nevertheless, one can associate an effective energy–momentum tensor with the gravitational field of a finite volume. Specifically, we can average over a spatial volume that is much larger than the wavelength of the relevant gravitational waves to obtain

$$t_{\mu\nu} = \frac{c^4}{8\pi G_\mathrm{N}}\left[\left\langle R^{(2)}_{\mu\nu}\right\rangle - \frac{1}{2}\eta_{\mu\nu}\left\langle R^{(2)}\right\rangle\right], \tag{13.46}$$

where $\langle\cdots\rangle$ stands for the average over many wave cycles.

Let us calculate the energy flux carried by a linearly polarized plane wave, say the h_+ state, propagating in the z direction. The metric and its inverse, accurate up to first order in perturbation, in the T T gauge can be written as

$$g_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1+\tilde{h}_+ & 0 & 0 \\ 0 & 0 & 1-\tilde{h}_+ & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad g^{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1-\tilde{h}_+ & 0 & 0 \\ 0 & 0 & 1+\tilde{h}_+ & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{13.47}$$

where

$$\tilde{h}_+ = h_+ \cos\left[\omega(t - z/c)\right]. \tag{13.48}$$

To obtain the energy–momentum tensor of the gravity wave by way of $R^{(2)}_{\mu\nu}$ as in (13.45), we need first to calculate the Christoffel symbols by differentiating the metric of (13.47). It can be shown (Problem 13.3) that the nonvanishing elements are

$$\Gamma^1_{10} = \Gamma^1_{01} = \Gamma^0_{11} = \frac{1}{2}(\partial_0\tilde{h}_+ - \tilde{h}_+\partial_0\tilde{h}_+) \tag{13.49}$$

and

$$\Gamma^1_{13} = \Gamma^1_{31} = -\Gamma^3_{11} = -\frac{1}{2}(\partial_0\tilde{h}_+ - \tilde{h}_+\partial_0\tilde{h}_+). \tag{13.50}$$

The Riemann tensor has the structure of $(\partial\Gamma + \Gamma\Gamma)$. Since we are interested in calculating only $O(h^2)$, the above $\tilde{h}_+\partial_0\tilde{h}_+$ factor in the Christoffel symbols can only enter in the $\partial\Gamma$ terms, leading to the time-averaged term of $\langle\tilde{h}_+\partial_0\tilde{h}_+\rangle \propto \langle\sin[2\omega(t - z/c)]\rangle = 0$. Hence we will drop the $\tilde{h}_+\partial_0\tilde{h}_+$ terms in (13.49) and (13.50), and calculate the (averaged) curvature tensor in (11.58) by dropping the $\langle\partial\Gamma\rangle$ factors,

$$\langle R^{(2)}_{\mu\nu}\rangle = \langle\Gamma^\alpha_{\alpha\lambda}\Gamma^\lambda_{\mu\nu} - \Gamma^\alpha_{\mu\lambda}\Gamma^\lambda_{\alpha\nu}\rangle. \tag{13.51}$$

A straightforward calculation (Problem 13.3) shows that

$$R^{(2)}_{11} = R^{(2)}_{22} = 0 \quad \text{and} \quad R^{(2)}_{00} = R^{(2)}_{33} = \frac{1}{2}(\partial_0\tilde{h}_+)^2 \tag{13.52}$$

leading to a vanishing Ricci scalar

$$R^{(2)} = \eta^{\mu\nu}R^{(2)}_{\mu\nu} = -R^{(2)}_{00} + R^{(2)}_{11} + R^{(2)}_{22} + R^{(2)}_{33} = 0. \tag{13.53}$$

In particular, the effective energy density of the gravitational plane wave in the plus polarization state as given by (13.46) and (13.52) is

$$t_{00} = \frac{c^4}{16\pi G_\mathrm{N}}\left\langle(\partial_0\tilde{h}_+)^2 + (\partial_0\tilde{h}_\times)^2\right\rangle, \tag{13.54}$$

where we have also added the corresponding contribution from the cross polarization state. If we choose to write the transverse traceless metric perturbation as $\tilde{h}_+ \equiv h^{\mathrm{TT}}_{11} = -h^{\mathrm{TT}}_{22}$ and $\tilde{h}_\times \equiv h^{\mathrm{TT}}_{12} = h^{\mathrm{TT}}_{21}$ and $h^{\mathrm{TT}}_{3i} = 0$ (with $i = 1, 2, 3$) as well as denote the time derivative $\dot{h}^{\mathrm{TT}}_{ij} \equiv \partial h^{\mathrm{TT}}_{ij}/\partial t$, we then have $\langle(\partial_0\tilde{h}_+)^2 + (\partial_0\tilde{h}_\times)^2\rangle = \frac{1}{2}\langle\dot{h}_{ij}^{\mathrm{TT}}\dot{h}_{ij}^{\mathrm{TT}}\rangle$. For a wave traveling at the speed c the energy flux, being related to the density by $f = ct_{00}$, hence can be expressed in terms of the metric perturbation as

$$f = \frac{c^3}{32\pi G_{\mathrm{N}}}\left\langle\dot{h}_{ij}^{\mathrm{TT}}\dot{h}_{ij}^{\mathrm{TT}}\right\rangle \tag{13.55}$$

with repeated indices summed over. It is useful to recall the counterpart in the more familiar electromagnetism. The EM flux is given by Poynting vector which is proportional to the product of time-derivatives of the vector potentials. Equation (13.55) shows that a gravitational wave is just the same, with the proportionality constant built out of c and G_{N}. One can easily check that c^3/G_{N} has just the right units (energy times time per unit area). It is a large quantity, again reflecting the stiffness of spacetime—a tiny disturbance in the metric corresponds to a large energy flux.

13.4.2 Emission of gravitational radiation

In the previous subsection we have expressed the energy flux of a gravitational wave in terms of the metric perturbation $h_{ij} = g_{ij} - \eta_{ij}$. Here we will relate h_{ij} to the source of gravitational wave by way of the linearized Einstein Eq. (13.22).

Calculate wave amplitude due to quadrupole moments

We shall be working in the long wavelength limit for a field-point far away from the source. Let D be the dimension of the source, this limit corresponds to

$r \gg D$	large distance from source,
$\lambda \gg D$	long wavelength.

The long wavelength $D\lambda^{-1} \sim D\omega c^{-1}$ approximation means a low velocity limit for the source particles. In such a limit we can approximate the integral over the energy–momentum source in (13.22) as

$$\int d^3\mathbf{x}'\frac{T_{\mu\nu}(\mathbf{x}', t - |\mathbf{x}-\mathbf{x}'|/c)}{|\mathbf{x}-\mathbf{x}'|} \longrightarrow \frac{1}{r}\int d^3\mathbf{x}' T_{\mu\nu}\left(\mathbf{x}', t - \frac{r}{c}\right)$$

because the harmonic source, in the long wave limit $T_{\mu\nu} \propto \cos[\omega t - (2\pi/\lambda) \times |\mathbf{x}-\mathbf{x}'|]$, will not change much when integrated over the source. To calculate the energy flux through (13.55) we have from (13.22)

$$h_{ij}(\mathbf{x}, t) = \frac{4G_{\mathrm{N}}}{c^4 r}\int d^3\mathbf{x}' T_{ij}\left(\mathbf{x}', t - \frac{r}{c}\right), \tag{13.56}$$

where we have not distinguished between h_{ij} and $\bar{h}_{ij}$ as they are the same in the TT gauge.

To calculate $\int d^3\mathbf{x}' T_{ij}(\mathbf{x}')$ we find it convenient to convert it into a second mass moment by way of the energy–momentum conservation relation $\partial_\mu T^{\mu\nu} = 0$.

Differentiating ∂_0 one more time, the equations $\partial_0\partial_\mu T^{\mu 0} = 0$ leads to

$$\frac{\partial^2 T^{00}}{c^2 \partial t^2} = -\frac{\partial^2 T^{i0}}{c\partial t \partial x^i} = -\frac{\partial}{\partial x^i}\frac{\partial T^{0i}}{c\partial t}.$$

We can apply the conservation relation $\partial T^{0i} + \partial_j T^{ij} = 0$ one more time to get

$$\frac{\partial^2 T^{00}}{c^2 \partial t^2} = +\frac{\partial^2 T^{ij}}{\partial x^i \partial x^j}.$$

Multiply both sides by $x^k x^l$ and integrate over the source volume:

$$\frac{\partial^2}{c^2 \partial t^2}\int d^3\mathbf{x} T^{00} x^k x^l = \int d^3\mathbf{x} \frac{\partial^2 T^{ij}}{\partial x^i \partial x^j} x^k x^l = 2\int d^3\mathbf{x} T^{kl}. \tag{13.57}$$

To reach the last equality we have performed two integrations-by-parts and discarded the surface terms because the source dimension is finite. Combining (13.56) and (13.57) we have

$$h_{ij}(\mathbf{x}, t) = \frac{2G_\mathrm{N}}{c^4 r}\ddot{I}_{ij}\left(t - \frac{r}{c}\right), \tag{13.58}$$

where I_{ij} is the second mass moment, after making the Newtonian approximation of $T_{00} = \rho c^2$ with $\rho(\mathbf{x})$ being the mass density, given by

$$I_{ij} = \int d^3\mathbf{x} \rho(\mathbf{x}) x_i x_j \tag{13.59}$$

and the double dots over I_{ij} indicate second-order time-derivatives.

We have already explained that, just as the electromagnetic case, there is no monopole radiation (Birkhoff's theorem). But unlike electromagnetism, there is also no gravitational dipole radiation because the time derivative of the dipole moment,

$$\ddot{\mathbf{d}} = \int d^3\mathbf{x} \rho(\mathbf{x}) \dot{\mathbf{v}} = 0, \tag{13.60}$$

which is the total force on the system, vanishes for an isolated system (reflecting momentum conservation). Thus, the leading gravitational radiation must be quadrupole radiation.

Summing over the flux in all directions in the T T gauge

Since our calculations are performed in the transverse traceless gauge, the relation in (13.58) suggests that the mass moment should have the same traceless and transverse structure as the metric perturbation h_{ij}^{TT}. We shall work with the "reduced mass moment" which is traceless:

$$\tilde{I}_{ij} = I_{ij} - \frac{1}{3}\delta_{ij} I_{kk}, \tag{13.61}$$

I_{kk} being the trace of I_{ij} and for a plane wave propagating in the z direction it has the form of (13.32)

$$\tilde{I}_{ij} = \begin{pmatrix} \tilde{I}_+ & \tilde{I}_\times & 0 \\ \tilde{I}_\times & -\tilde{I}_+ & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

The combination of fields in (13.55) and (13.54) implies that we calculate the sum

$$\tilde{I}_+^2 + \tilde{I}_\times^2 = \frac{1}{4}(\tilde{I}_{11} - \tilde{I}_{22})^2 + \tilde{I}_{12}^2. \tag{13.62}$$

For the calculation to be performed below we need to rewrite (Problem 13.4) this result explicitly as the mass moment for a wave propagating in the z direction as

$$[\tilde{I}_+^2 + \tilde{I}_\times^2]_z = \frac{1}{4}[2\tilde{I}_{ij}\,\tilde{I}_{ij} - 4\tilde{I}_{i3}\,\tilde{I}_{i3} + \tilde{I}_{33}\,\tilde{I}_{33}], \tag{13.63}$$

where we have used $\tilde{I}_{33} = -\tilde{I}_{11} - \tilde{I}_{22}$ because the reduced moment is traceless.

To calculate the total power emitted by the source, we need to integrate over the flux for a wave propagating out in **all** directions. We need to generalize the result in (13.63) to moments for a plane wave propagating in an arbitrary direction, specified by the unit vector $\mathbf{n}$:

$$[\tilde{I}_+^2 + \tilde{I}_\times^2]_n = \frac{1}{4}[2\tilde{I}_{ij}\,\tilde{I}_{ij} - 4\tilde{I}_{ik}\,\tilde{I}_{il}\,n_k\,n_l + \tilde{I}_{ij}\,\tilde{I}_{kl}\,n_i\,n_j\,n_k\,n_l]. \tag{13.64}$$

Integrating over all directions, we obtain

$$\begin{aligned}\int \frac{1}{4}&[2\tilde{I}_{ij}\,\tilde{I}_{ij} - 4\tilde{I}_{ik}\,\tilde{I}_{il}\,n_k\,n_l + \tilde{I}_{ij}\,\tilde{I}_{kl}\,n_i\,n_j\,n_k\,n_l]d\Omega \\ &= \pi\left(2 - \frac{4}{3} + \frac{2}{15}\right)\tilde{I}_{ij}\,\tilde{I}_{ij} = \frac{4\pi}{5}\tilde{I}_{ij}\,\tilde{I}_{ij}\end{aligned} \tag{13.65}$$

after using the formulae

$$\int d\Omega = 4\pi,$$

$$\int n_k n_l d\Omega = \frac{4\pi}{3}\delta_{kl},$$

$$\int n_i n_j n_k n_l d\Omega = \frac{4\pi}{15}(\delta_{ij}\delta_{kl} + \delta_{kj}\delta_{il} + \delta_{ik}\delta_{jl}).$$

These integration results are easy to understand: the only available symmetric tensor that is invariant under rotation is the Kronecker delta δ_{ij}. After fixing the tensor structure of the integrals, the coefficients in front, $4\pi/3$ and $4\pi/15$, can be obtained by contracting the indices on both sides and using the relation $\delta_{ij}\delta_{ij} = 3$.

Integrating the flux (13.55) over all directions by using the result of (13.65) we arrive at the expression for the total luminosity

$$\frac{dE}{dt} = \int f \cdot r^2 d\Omega = \frac{G_\mathrm{N}}{5c^5}\left\langle \dddot{\tilde{I}}{}_{ij}^{\mathrm{TT}}\,\dddot{\tilde{I}}{}_{ij}^{\mathrm{TT}} \right\rangle. \tag{13.66}$$

Let us recapitulate: the energy carried away by gravitational waves must be proportional to the square of the time-derivative of the wave amplitude (recall the Poynting vector), which is the second derivative of the quadrupole moment (cf. (13.58)). The energy flux falls off like r^{-2}. To get the total luminosity by integrating over a sphere of radius r, the dependence of radial distance disappears. The factor of $G_\mathrm{N}c^{-5}$ must be present on dimensional grounds. The detailed calculation fixes the proportional constant of $\frac{1}{5}$ and we have the gravitational wave luminosity in the quadrupole approximation displayed above.

13.4.3 Binary pulsar PSR 1913+16

A radio survey, using the Arecibo Radio Telescope in Puerto Rico (Fig. 13.4), for pulsars in our galaxy made by Russel Hulse and Joseph Taylor discovered

Fig. 13.4 The Arecibo Radio telescope.

the unusual system PSR 1913+16. Observations made since 1974 allowed them to check GR to great precision including the verification for the existence of gravitational waves as predicted in Einstein's theory.

From the small changes in the arrival times of the pulses recorded in the past decades a wealth of the properties of this binary system can be extracted. This is achieved by modeling the orbit dynamics and expressing these in terms of the arrival time of the pulse. Different physical phenomena (such as bending of the light, periastron advance, etc.) are related to the pulse time through different combinations of system parameters. In this way the masses and separation of the stars and the inclination and eccentricity of their orbit can all be deduced. In the following we present a recent compilation given by Weisberg and Taylor (2003):

$$\begin{array}{ll}
\text{pulsar mass} & M_\mathrm{p} = 1.4408 \pm 0.0003\, M_\odot, \\
\text{companion mass} & M_\mathrm{c} = 1.3873 \pm 0.0003\, M_\odot, \\
\text{eccentricity} & e = 0.6171338 \pm 0.000004, \\
\text{binary orbit period} & P_\mathrm{b} = 0.322997462727 \text{ d}, \\
\text{orbit decay rate} & \dot{P}_\mathrm{b} = (-2.4211 \pm 0.0014) \times 10^{-12} \text{ s/s}.
\end{array} \tag{13.67}$$

It is interesting to note that these two neutron stars have just the masses $1.4M_\odot$ of the Chandrasekhar limit.

In this section, we shall demonstrate that from these numbers, without any adjustable parameters, we can compute the decrease (decay) of orbit period due to gravitational radiation by the orbiting binary system. Instead of a full scale GR calculation, we shall consider the simplified case of two equal mass stars in a circular orbit (Fig. 13.5), as all essential features of gravitational radiation and orbit decay can be easily calculated. At the end we then quote the exact result when $M_\mathrm{p} \neq M_\mathrm{c}$ in an orbit with high eccentricity as a straightforward modification of the result obtained by our simplified calculation.

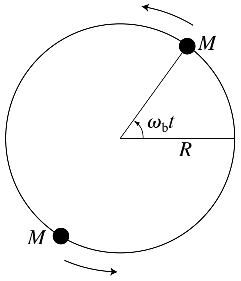

Fig. 13.5 Two equal masses circulating each other in a circular orbit with angular frequency of ω_b.

Energy loss due to gravitational radiation

Let us first concentrate on the instantaneous position of one of the binary stars as shown in Fig. 13.5:

$$x_1(t) = R\cos\omega_\mathrm{b}t, \quad x_2(t) = R\sin\omega_\mathrm{b}t, \quad x_3(t) = 0.$$

From this we can calculate the second mass moment according to (13.59),

$$\begin{aligned} I_{11} &= 2MR^2\cos^2\omega_\mathrm{b}t, \\ I_{22} &= 2MR^2\sin^2\omega_\mathrm{b}t, \\ I_{12} &= 2MR^2\sin\omega_\mathrm{b}t\cos\omega_\mathrm{b}t, \end{aligned}$$

leading to the traceless reduced moment as defined in (13.61),

$$\tilde{I}_{ab} = I_{ab} - \frac{1}{2}\delta_{ab}I_{cc} = I_{ab} - MR^2\delta_{ab},$$

so that

$$\begin{aligned} \tilde{I}_{11} &= MR^2\cos 2\omega_\mathrm{b}t, \\ \tilde{I}_{22} &= -MR^2\cos 2\omega_\mathrm{b}t, \\ \tilde{I}_{12} &= MR^2\sin 2\omega_\mathrm{b}t. \end{aligned}$$

The quadrupole formula (13.66) for luminosity involves time derivatives. For the simple sinusoidal dependence given above, each derivative just brings down a factor of $2\omega_\mathrm{b}$; together with the averages $\langle\sin^2\rangle = \langle\cos^2\rangle = \frac{1}{2}$, we obtain the rate of energy loss:

$$\frac{dE}{dt} = \frac{G_\mathrm{N}}{5c^5}(2\omega_\mathrm{b})^6\langle\tilde{I}_{11}^2 + \tilde{I}_{22}^2 + 2\tilde{I}_{12}^2\rangle = \frac{128G_\mathrm{N}}{5c^5}\omega_\mathrm{b}^6M^2R^4. \tag{13.68}$$

From energy loss to orbit decay

Energy loss through gravitational radiation leads to orbit decay, namely the decrease in orbit period P_b of the binary system. We start the calculation of this orbit period change through the relation $(dP_\mathrm{b})/P_\mathrm{b} \propto -(dE)/E$. Again we shall only work out the simpler situation of a binary pair of equal mass M separated by $2R$ in circular motion. The total energy being

$$E = MV^2 - \frac{G_\mathrm{N}M^2}{2R} \tag{13.69}$$

with velocity determined by Newtonian equation of motion

$$M\frac{V^2}{R} = \frac{G_\mathrm{N}M^2}{(2R)^2}$$

or

$$V^2 = \frac{G_\mathrm{N}M}{4R}, \tag{13.70}$$

so that the total energy of the binary system (13.69) comes out to be

$$E = -\frac{G_\mathrm{N}M^2}{4R}. \tag{13.71}$$

We wish to have an expression of the energy in terms of the orbit period by replacing R using (13.70):

$$R = \frac{G_N M}{4V^2} = \frac{G_N M}{4}\left(\frac{2\pi R}{P_b}\right)^{-2} \quad \text{or} \quad R^3 = \frac{G_N M}{16\pi^2} P_b^2. \tag{13.72}$$

Plugging this back into (13.71), we have

$$E = -M\left(\frac{\pi M G_N}{2}\right)^{2/3} P_b^{-2/3}. \tag{13.73}$$

Through the relation, $dE/E = -\frac{2}{3}dP_b/P_b$, so that the rate of period-decrease $\dot{P}_b \equiv dP_b/dt$ can be related to the energy loss rate:

$$\dot{P}_b = -\frac{3P_b}{2E}\left(\frac{dE}{dt}\right). \tag{13.74}$$

Substituting in the expression (13.73) for E, (13.68) for dE/dt where the wave frequency is given by the orbit frequency $\omega_b = 2\pi/P_b$ and where R is given by (13.72), we have

$$\dot{P}_b = -\frac{48\pi}{5c^5}\left(\frac{4\pi G_N M}{P_b}\right)^{5/3}. \tag{13.75}$$

That the orbit for PSR 1913+16, rather than circular, is elliptical with high eccentricity can be taken into account (Peters and Mathews, 1963) with the result involving a multiplicative factor of

$$\frac{1 + (73/24)e^2 + (37/96)e^4}{(1 - e^2)^{7/2}} = 11.85681, \tag{13.76}$$

where we have used the observed binary orbit eccentricity as given in (13.67). That the pulsar and its companion have slightly different masses, $M_p \neq M_c$

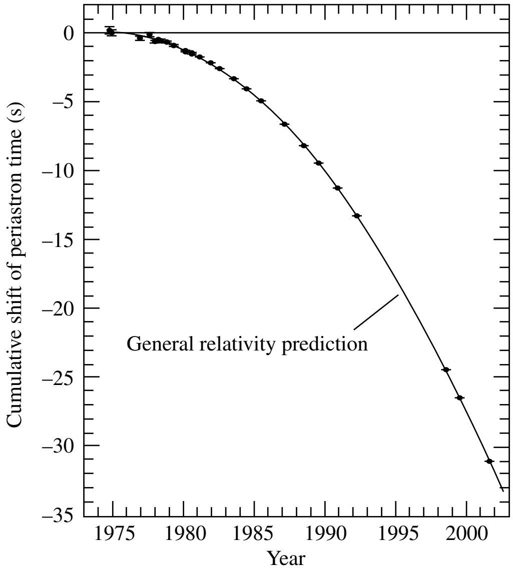

Fig. 13.6 Gravitational radiation damping causes orbit decay of the binary pulsar PSR 1913+16. Plotted in Weisberg and Taylor (2003) is the accumulating shift in the epoch of periastron (the point of closest approach between the pulsar and its companion star). The parabola is the GR prediction, and observations are depicted by data points. In most cases the measurement uncertainties are smaller than the line widths. The data gap in the 1990s reflects the downtime when the Arecibo observatory was being upgraded.

means we need to make the replacement $(2M)^{5/3} \longrightarrow 4M_\mathrm{p}M_\mathrm{c}(M_\mathrm{p}+M_\mathrm{c})^{-1/3}$. The exact GR prediction is found to be

$$\begin{aligned}\dot{P}_{\mathrm{b\,GR}} &= \frac{-192\pi M_\mathrm{p}M_\mathrm{c}}{5c^5(M_\mathrm{p}+M_\mathrm{c})^{1/3}} \frac{1+(73/24)e^2+(37/96)e^4}{(1-e^2)^{7/2}} \left(\frac{2\pi G_\mathrm{N}}{P_\mathrm{b}}\right)^{5/3} \\ &= -(2.40247 \pm 0.00002) \times 10^{-12}\ \mathrm{s/s}. \end{aligned} \tag{13.77}$$

This is to be compared to the observed value corrected for the galactic acceleration of the binary system and the sun, which also causes a change of orbit period $\dot{P}_{\mathrm{b\,gal}} = -(0.0125 \pm 0.0050) \times 10^{-12}$ s/s. From the measured values given in (13.67), we then have

$$\begin{aligned}\dot{P}_{\mathrm{b\,corrected}} &= \dot{P}_{\mathrm{b\,observed}} - \dot{P}_{\mathrm{b\,gal}} \\ &= -(2.4086 \pm 0.0052) \times 10^{-12}\ \mathrm{s/s} \end{aligned} \tag{13.78}$$

in excellent agreement with the theoretical prediction shown in (13.77). This result (Fig. 13.6) provides strong confirmation of the existence of gravitational radiation as predicted by Einstein's theory of GR.

With the confirmation of the existence of gravitational radiation according to Einstein's general theory of relativity, the next stage will be the detection of gravitational waves through interferometer observations to confirm the expected wave kinematics, and tests of various strong field situations. But just like all pioneering efforts of fresh ways to observe the universe, gravitational wave observatories will surely discover new phenomena that will deepen and challenge our understanding of astronomy, gravitation, and cosmology.

Review questions

1. Give a qualitative discussion showing why one would expect gravitational waves from Einstein's GR theory of gravitation, but not from Newton's theory.
2. Why is it important to have a gravitational wave observatory?
3. What approximation is made to have the linearized theory of GR? In this framework, how should we view the propagation of gravitational waves?
4. What are the differences and similarities between electromagnetic and gravitational waves?
5. What is a gauge transformation in the linearized theory? What is the Lorentz gauge? Can we make further gauge transformations within the Lorentz gauge?
6. Consider a set of test particles, all of them lying in a circle except one at the center. When a gravitational wave with the + polarization passes through them what will be the relative displacements of these particles going through one period of the wave? How would the relative displacement be different if the polarization is of the × type?
7. Give a qualitative argument showing that the wave strain is of the order $\epsilon G_\mathrm{N}M/rc^2$ where ϵ is a relativistic correction factor typically less than unity. Such a strain would be $O(10^{-21})$ when the wave is generated by a solar mass source in the Virgo cluster ($r \approx 15$ Mpc) from us.
8. Using what you know of the Poynting vector as the energy flux of an EM wave, guess the form of energy flux in terms of the gravitational wave amplitude. What should be the proportionality constant (up to some numerical constant that can only be derived by detailed calculation)?
9. The leading term in gravitational radiation is quadrupole. Why is there no monopole and dipole radiation?
10. PSR 1913+16 is a binary pulsar system. What is a pulsar? What is being observed? Which results show strong evidence for the existence of gravitation waves as predicted by GR?

Problems

(13.1) **Gauge transformations**

(a) Show that the gauge transformation for the trace-reversed perturbation $\bar{h}_{\mu\nu}$ in (13.20) follows from (13.17) and (13.19).

(b) Demonstrate the existence of the Lorentz gauge by showing that, starting with an arbitrary coordinate system where $\partial^{\mu}\bar{h}_{\mu\nu} \neq 0$, one can always find a new system such that $\partial^{\mu}\bar{h}'_{\mu\nu} = 0$ with a gauge vector function χ_{μ} being the solution to the inhomogeneous wave equation $\Box\chi_{\nu} = \partial^{\mu}\bar{h}_{\mu\nu}$. This also means that one can make further coordinate transformations within the Lorentz gauge, as long as the associated gauge vector function satisfies the wave equation

$$\Box\chi_{\nu} = 0. \qquad (13.79)$$

(c) The solution to Eq. (13.79) may be written as $\chi_{\nu} = X_{\nu}e^{ikx}$ where k^{α} is a null-vector. Show that the four constants X_{ν} can be chosen such that the polarization tensor in the metric perturbation $h_{\mu\nu}(x)$ is traceless $\epsilon^{\mu}_{\mu} = 0$ and every 0th component vanishes $\epsilon_{\mu 0} = 0$.

(13.2) **Wave effect via the deviation equation** As we have shown in Section 13.3.1, a gravitational wave can only be detected through the tidal effect. Since the equation of geodesic deviation is an efficient description of the tidal force, show that the results of (13.37) and (13.38) can be obtained by using this equation.

(13.3) **$\Gamma^{\mu}_{\nu\lambda}$ and $R^{(2)}_{\mu\nu}$ in the TT gauge** Show that the Christoffel symbols of (13.49) and (13.50), as well as the second-order Ricci tensor (13.52), are obtained in the T T gauge with the metric given in (13.47).

(13.4) **Checking the equivalence of (13.62) and (13.63)** Show that

$$(\tilde{I}_{11} - \tilde{I}_{22})^2 + 4\tilde{I}_{12}^2 = 2\tilde{I}_{ij}\tilde{I}_{ij} - 4\tilde{I}_{i3}\tilde{I}_{i3} + \tilde{I}_{33}\tilde{I}_{33}.$$

This page intentionally left blank

Supplementary notes

A

- These appendices contain the supplementary material for various sections in the main text (marked by the bracket in the headings).

A.1 The twin paradox (Section 2.3.4)

The well-known “twin paradox” is an instructive example that sheds light on several basic concepts in relativity. We shall work it out in detail, and arrive at the resolution of the paradox (as a reciprocity puzzle) when realizing that non-inertial frames are involved. (See, for example, Ellis and Williams, 1988.)

The paradox of a twin’s asymmetric aging

Two siblings, Al and Bill, are born on the same day. Al goes on a long journey at high speed in a spaceship; Bill stays at home. The biological clock of Al will be measured by the stay-at-home Bill to run slow. When Al returns he should be younger than Bill.

A definite example with numbers

For simplicity, let us consider a definite case where Al travels outward at $\beta = \frac{4}{5}$ for 15 years, then returns (i.e. coming inward) at $\beta = -\frac{4}{5}$ for 15 years. Both periods of 15 years are measured in Al’s rocketship; and $\gamma = \frac{5}{3}$ for both $\beta = \pm\frac{4}{5}$. When A and B meet again, Al should be younger: while A has aged 30 years, B would have aged $\frac{5}{3} \times 30 = 50$ years according to SR time dilation:

$$\text{A vs. B} = 30 \text{ vs. } 50 \qquad \text{(B's viewpoint)}. \tag{A.1}$$

Of course, this SR prediction of asymmetric aging of the twins, while counter-intuitive according to our low-velocity experience, is nothing “paradoxical”—just an example of time-dilation, which is counter-intuitive, but true. The “twin paradox” is just a more dramatic way of saying that any two travelers A and B would find their watches no longer synchronized when they meet again after a journey following two separate routes. It is easy to understand this because their separate worldlines will have different spacetime lengths, which are readings of their respective proper times. These quantities being invariants (independent of the coordinate frames), we can choose to calculate them in their respective rest frames: Al’s proper time is $15 + 15 = 30$, while Bill’s proper time is 50 (see Fig. A.1). Nevertheless, it is worthwhile to work through some details as it will provide us with further insight as to the nature of time coordinate changes when we change reference frames. Before doing this, we will state below a question which is somewhat more "paradoxical."

The twin paradox as a reciprocity puzzle

If relativity is truly relative, we could just as well consider this separation and reunion from the viewpoint of Al, who sees Bill as in a moving frame. So, when Bill "returns," it is Bill who should have stayed younger. From the viewpoint of A, Bill should be younger: while A has aged 30 years, B would have aged $(\frac{5}{3})^{-1} \times 30 = 18$ years.

$$\text{A vs. B} = 30 \text{ vs. } 18 \qquad \text{(A's viewpoint).} \tag{A.2}$$

Thus Bill's age has been found in one case, Eq. (A.1), to be 50 and in another case 18—a full 32 years difference. Which viewpoint, which theory, is the correct one?

Checking theories by measurements

To answer this question, we can carry out the following measurement of Bill's age. Let the stay-at-home Bill celebrate his birthdays by setting off firework displays, which Al, with a powerful telescope on board his spaceship, can always observe. So the theory can be checked with experiments by the number of firework flashes Al sees during his 30-year long journey. If he sees 18 flashes A's viewpoint is right, if 50, then B is right.

Let us carry out this measurement:

During the outward-bound journey. Al sees a flash at every $\Delta t_{\rm A}$ interval, which differs from $\Delta t_{\rm B} = 1$ year: first of all there is the time dilation effect $\gamma \Delta t_{\rm B}$, but there is also the fact that between the flashes A and B have increased their separation by an amount of $v\Delta t_{\rm B}$. Therefore, to reach Al in the space-ship, the light signal (i.e. the flashes) has to take an extra interval of $v\Delta t_{\rm B}/c$, which also has to be dilated by a factor of γ. (This is a particular realization of the nonsynchronicity of clocks as discussed in reference to Fig. 2.14. Keep in mind that the fireworks act as clocks in this situation.)

$$\Delta t_{\rm A}^{(\rm out)} = \gamma(1+\beta)\Delta t_{\rm B} = \frac{5}{3}\left(1+\frac{4}{5}\right) = 3 \text{ years.}$$

Namely, during the 15-years' outward bound journey, Al sees Bill's flashes every 3 years, thus a total of 5 flashes.

During the inward-bound journey. β reverses sign, hence Al sees a flash at an interval of[1]

$$\Delta t_{\rm A}^{(\rm in)} = \frac{5}{3}\left(1-\frac{4}{5}\right) = \frac{1}{3} \text{ year,}$$

and thus altogether 45 flashes.

In this way, Al sees a total of 50 flashes. This proves that B's viewpoint is correct: while the traveling twin Al has aged 30 years, the stay-at-home twin has aged 50 years.

[1] We can also understand the difference in the observed frequency of birthday fireworks by the Doppler formula. The relative velocities for the outward and inward bound trips are $\beta = \pm\frac{4}{5}$. Equation (10.48) yields $\omega' = 3\omega$, and $\omega' = \omega/3$, respectively. This is just the frequency changes of birthday fireworks observed.

Reciprocity nature of relativistic effects

Applying the basic formula for time dilation, we see that Al observes Bill's clock to run slow:

$$t_{\mathrm{B}} = \gamma t_{\mathrm{A}}, \tag{A.3}$$

while Bill observes Al's clock to run slow:

$$t_{\mathrm{A}} = \gamma t_{\mathrm{B}}. \tag{A.4}$$

Each observer sees the other's clock to run slow. Thus we can conclude "relativity is truly relative." But, there is an apparent contradiction: "which set of time intervals is actually longer?!" The resolution is based on the realization that two different types of time are being compared. Failure to distinguish them have led to the confusion. Equation (A.3) is from Al's view point, and Eq. (A.4) is from Bill's view point. Thus to be precise we should have labeled the times in these two equations differently:

$$t_{\mathrm{B}}^{(\mathrm{A})} = \gamma t_{\mathrm{A}}^{(\mathrm{A})}, \quad \text{and} \quad t_{\mathrm{A}}^{(\mathrm{B})} = \gamma t_{\mathrm{B}}^{(\mathrm{B})}. \tag{A.5}$$

and $t_{\mathrm{A,B}}^{(\mathrm{A})} \neq t_{\mathrm{A,B}}^{(\mathrm{B})}$. In fact since $t_{\mathrm{A}}^{(\mathrm{A})}$ and $t_{\mathrm{B}}^{(\mathrm{B})}$ are the respective proper times of A and B, we have the usual time dilation relations

$$t_{\mathrm{A}}^{(\mathrm{B})} = \gamma t_{\mathrm{A}}^{(\mathrm{A})}, \quad \text{and} \quad t_{\mathrm{B}}^{(\mathrm{A})} = \gamma t_{\mathrm{B}}^{(\mathrm{B})}. \tag{A.6}$$

From this, it is straightforward to check the two relations in (A.5) are entirely compatible with each other.

In connection with this discussion of the reciprocity relation, it will be useful to work through another measurement when it is the traveling Al who celebrates his birthdays by sending out light signals. In this arrangement, Al's yearly flashes are received every 3 years by Bill at home during the outward bound journey, thus a total 45 years before Al turns around. Thereafter, the flashes are received every 4 months, thus a total 15 flashes in 5 years, hence 30 flashes in 50 years of Bill's time. This agrees with the above measurements. This example shows explicitly the consistency of the result, even though each observer sees the other's clock to run slow.

The reciprocity puzzle resolved

Although we properly recognize that two types of time measurements are involved (hence resolving the reciprocity puzzle), we still have the question of why B's viewpoint is correct, and not A's? The reciprocity relation is not applicable here: while B stays as an inertial frame observer throughout the journey, A cannot be—he must turn around! Thus A's viewpoint cannot be represented by a single inertial frame of reference. We have to use at least two inertial frames to describe A's trip. The turn-around must necessarily involve acceleration to go from one to the other inertial frame. Non-inertial frames must be invoked and this goes beyond special relativity (SR).

In the above, we have considered the simplest possible case having only the minimum number of inertial frames: O = rest frame of B, and O′ and O″ = rest frames of A during his outward and inward bound segments of the journey. They are represented by three straight worldlines in the spacetime diagram, plotted from the viewpoint of B, in Fig. A.1. The OP interval corresponds to the duration

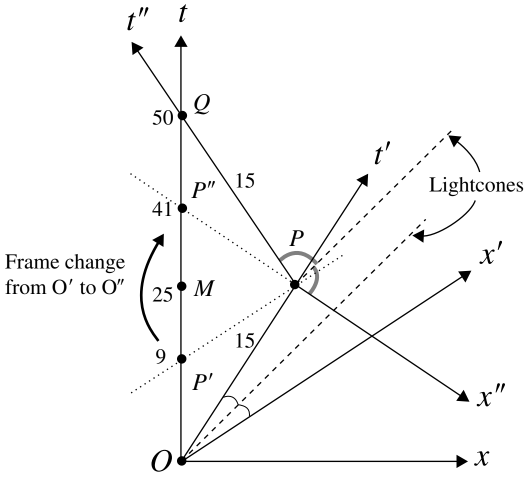

Fig. A.1 Three worldlines of the twin paradox: OQ is that for the stay-at-home Bill, OP that for the outward-bound part ($\beta = \frac{4}{5}$), PQ that for the inward-bound part ($\beta = -\frac{4}{5}$) of the Al's journey. M is the midpoint between O and Q. These three lines define three inertial frames: O, O′, and O″ systems. When Al changes from the O′ to the O″ system at P the point that is simultaneous (with P) along Bill's worldline OQ jumps from point P′ to P″. From the viewpoint of Bill, this is a leap of 32 years.

of $\Delta t' = 15$ years, while the PQ interval to $\Delta t'' = 15$ years. Let us examine this journey from the viewpoints of these three inertial frames.

1. The system O (the viewpoint of stay-at-home Bill): the worldpoints P and M are simultaneous.

$$\mathrm{OM} = \gamma \Delta t' = \frac{5}{3} \times 15 = 25,$$

$$\mathrm{MQ} = \gamma \Delta t'' = \frac{5}{3} \times 15 = 25.$$

Adding up, Bill ages a total of OQ = 50 years.

2. The system O′ (the viewpoint of outward bound Al): the worldpoints P and P′ are simultaneous so that $\Delta t'(\mathrm{P'}) = \Delta t'(\mathrm{P}) = 15$, with P′ being stationary in the O frame, $\Delta x(\mathrm{P'}) = 0$. The Lorentz transformation states:

$$\Delta t'(\mathrm{P'}) = \gamma \left[\Delta t(\mathrm{P'}) - \beta \Delta x(\mathrm{P'})/c\right] = \gamma \Delta t(\mathrm{P'}).$$

Thus

$$\mathrm{OP'} = \Delta t(\mathrm{P'}) = \gamma^{-1} \Delta t'(\mathrm{P'}) = \frac{3}{5} \times 15 = 9.$$

3. The system O″ (the viewpoint of inward bound Al): the worldpoints P and P″ are simultaneous. We can similarly obtain

$$\mathrm{QP''} = \gamma^{-1} \Delta t'' = \frac{3}{5} \times 15 = 9.$$

Thus OP″ represents an interval of 50 − 9 = 41—as compared to OP′ being 9.

Namely, **an instant before** the turning point P, the two points P and P′ are (with respect to the spaceship) simultaneous; **an instant after** P (after the ship has turned around) it is P and P″ that are simultaneous. But P′ has its time coordinate in the O frame as $t(\mathrm{P'}) = 9$ years; the worldpoint P″ is viewed to have the $t(\mathrm{P''}) = 41$ years.

What we have learned here emphasizes again the point that time is just another coordinate label. When we change the frame of reference, all coordinates (x, y, z, t) make their corresponding changes. Two points P′ and P″, being respectively 9 and 41 as measured in the O system are simultaneous to P when viewed from two different inertial frames, the O′ and O″ systems, respectively. Thus a difference of 32 years is brought about simply by a change of coordinate: $\text{O}' \longrightarrow \text{O}''$.

Remark: We can also understand this difference formally as a time dilation effect between the two inertial frames of O′ and O″. Let us first find out the relative velocity $\bar{\beta}$ between these two frames by the velocity addition rule of (2.24) for $\beta' = \frac{4}{5}$ and $\beta'' = -\frac{4}{5}$:

$$\bar{\beta} = \frac{\beta' - \beta''}{1 - \beta'\beta''} = \frac{40}{41}. \tag{A.7}$$

This translates into a gamma factor $\bar{\gamma} = \frac{41}{9}$. The time dilation effect between the O′ and O″ frames is such that a time interval of $t(\text{P}') = 9$ years in O′ is viewed as $t(\text{P}'') = \bar{\gamma}t(\text{P}') = 41$ years in the O″ frame.

This accounts for the reason why Al, without taking this missed 32 years into the calculation, erroneously concluded that Bill had only aged 18 years.

This is as far as SR can go. To understand the physical origin of this extra 32 years we need to consider accelerating frames of reference. As we find in Chapter 3, accelerating frames are equivalent to inertial frames with gravity, this extra 32 years is due to the **gravitational time-dilation** effect, see Section 3.3.1. Thus, in principle, we need to go to general relativity (GR) (see Problem 3.3) in order to completely resolve the twin paradox.

A.2 A glimpse of advanced topics in black hole physics (Section 6.4)

In this Appendix, we shall offer some brief remarks on the results that have been discovered about black holes, beyond the simplest non-rotating spherical case discussed in Section 6.4. Any detailed discussion of these advanced topics is beyond the scope of this introductory exposition. Our purpose here is merely to alert the readers to the existence of a vast body of knowledge on topics such as rotating and charged black holes, nonspherical gravitational collapse, black hole thermodynamics, Hawking radiation, and quantum gravity, etc.

Beyond Schwarzschild black holes

- The Schwarzschild black hole is characterized by a single parameter, the stellar mass M. The GR solutions for the warped spacetime outside rotating and electrically charged mass sources are also known: they are, respectively, the **Kerr geometry** and the **Reissner–Nordström geometry**. This set of solutions is characterized by at most three parameters: the total mass M, angular momentum J, and electric charge Q. The lack of any detailed feature on a black hole has been described as "**black holes have no hair.**"

- The **singularity theorem** of GR states that any gravitational collapse that has proceeded far enough along its destiny will end in a physical singularity. Thus the $r = 0$ singularity encountered in the geometry outside a spherical source is not a peculiar feature of the spherical coordinate system.
- It is conjectured, and has not been proven in generality, that for the general nonspherical symmetric collapse GR also predicts the formation of an event horizon, shielding the physical singularity from all outside observers. This is called the **cosmic censorship conjecture**.

Black holes and quantum gravity

GR, as a classical field theory, is the $\hbar \to 0$ limit of the quantum theory of gravity. Putting it another way, quantum gravity, as the quantum description of space, time, and the universe, represents the union of GR with quantum mechanics.[2] All indications are that it is **the** fundamental theory of physics because its candidate theories (such as superstring theory) also encompass the description of strong, weak, and electromagnetic interactions. Thus, it is a theory of quantum gravity and unification. Although impressive advances have been made, this program is still very much a work-in-progress. It represents a major forefront in current theoretical physics research.

The Planck scale This is the scale at which physics must be described by quantum gravity. Soon after the 1900 discovery of **Planck's constant** in fitting the blackbody spectrum, Planck noted that a self-contained unit system of **mass–length–time** can be obtained from various combinations of Newton's constant G_N (gravity), Planck's constant $\hbar$ (quantum theory), and the velocity of light c (relativity). When we recall that $G_\mathrm{N}\cdot$ (**mass**)$^2 \cdot$ (**length**)$^{-1}$ has the unit of **energy**, and the natural scale of **energy·length** in relativistic quantum theory is $\hbar c$, we have the natural mass scale for quantum gravity, the **Planck mass**,

$$M_\mathrm{Pl} = \left(\frac{\hbar c}{G_\mathrm{N}}\right)^{1/2}. \tag{A.8}$$

From this we can immediately derive the other Planck scales:

$$\begin{aligned}
&\textit{Planck energy} && E_\mathrm{Pl} = M_\mathrm{Pl}c^2 = \left(\frac{\hbar c^5}{G_\mathrm{N}}\right)^{1/2} = 1.22 \times 10^{19}\ \mathrm{GeV} \\
&\textit{Planck length} && l_\mathrm{Pl} = \frac{\hbar c}{E_\mathrm{Pl}} = \left(\frac{\hbar G_\mathrm{N}}{c^3}\right)^{1/2} = 1.62 \times 10^{-33}\ \mathrm{cm} \\
&\textit{Planck time} && t_\mathrm{Pl} = \frac{l_\mathrm{Pl}}{c} = \left(\frac{\hbar G_\mathrm{N}}{c^5}\right)^{1/2} = 5.39 \times 10^{-44}\ \mathrm{s}
\end{aligned} \tag{A.9}$$

Such extreme scales are vastly beyond the reach of any laboratory setups. (Recall that the rest energy of a nucleon is about 1 GeV, and the highest energy the current generation of accelerators can reach is about 10^3 GeV.) The natural phenomena that can reach such an extreme density of $M_\mathrm{Pl}/(l_\mathrm{Pl})^3 = c^5/(\hbar G_\mathrm{N}) = 5.16\times10^{96}$ g/cm^3 are the physical singularities in GR: end points of gravitational collapse hidden inside a black hole horizon and the origin of the cosmological big bang. It is expected that quantum gravity will modify such GR singularity features.

[2] In the same sense a quantum field theory (e.g. quantum electrodynamics) is an union of SR (as embodied in a classical field theory such as Maxwell's electrodynamics) and quantum mechanics.

Hawking radiation The black hole physics of GR is a classical macroscopic description. For a microscopic theory we would need quantum gravity. We are familiar with the macroscopic physics as given by thermodynamics, which is obtained by averaging over atomic motions. While we do not have a fully developed theory of quantum gravity, one is curious to see whether the various candidate theories of quantum gravity have the correct features that can be checked by this micro–macro connection. It turns out that the blackbody radiation emitted by black holes, the Hawking radiation, provides such a handle.

The surprising theoretical discovery by Stephen Hawking that a black hole can radiate (contrary to the general expectation that nothing can come out of a black hole) was made in the context of a quantum description of particle fields in the background Schwarzschild geometry. Namely, the theoretical framework involves only a partial unification of gravity with quantum theory: while the fields of photons, electrons, etc., are treated as quantized fields (uniting SR and quantum theory), gravity is still described by the classical (non-quantum) theory of GR.

The quantum uncertainty principle of energy and time, $\triangle E \triangle t \gtrsim \hbar/2$, implies that processes violating energy conservation can occur, provided they take place in a sufficiently short time interval $\triangle t$. Such quantum fluctuations cause the empty space to become a medium with particle and antiparticle pairs appearing and disappearing. In normal circumstances such energy nonconserving processes cannot survive in the classical limit. (Hence the temporarily created and destroyed particles are called “virtual particles”.) However, if such random quantum fluctuations take place near an event horizon of a black hole, the virtual particles can become real because in such a situation energy conservation can be maintained on the macroscopic timescale. To understand this we need to take a deeper look at energy conservation. Conservation laws are usually associated with some symmetry in physics. Energy conservation reflects the invariance of physics laws under displacement in the time coordinate. This can be expressed in terms of invariance of the scalar product $p^\mu g_{\mu\nu}\xi^\nu$, where $p^\mu = (E, \mathbf{p}c)$ is the 4-momentum (cf. Section 10.1.2, also Problem 5.2), $g_{\mu\nu}$ the metric tensor, ξ^ν (called a Killing vector[3]) singles out the invariance direction—in this case, $\xi^\mu = (1, 0, 0, 0)$. Thus, for a quantum fluctuation from the vacuum into pair creation of a particle and antiparticle (with momenta p^μ and $\tilde{p}^\mu$, respectively), we must have $0 = p^\mu g_{\mu\nu}\xi^\nu + \tilde{p}^\mu g_{\mu\nu}\xi^\nu$. In a flat spacetime geometry, this relation takes on the familiar form of energy conservation: $0 = E + \tilde{E}$, which cannot be satisfied because both E and $\tilde{E}$ are positive. On the other hand, if such a fluctuation takes place near the black hole horizon and one particle travels across the event horizon (during the short time $\triangle t$), we have one particle outside the horizon $r > r^*$ and another inside $r < r^*$. According to Eq. (6.17), the Schwarzschild metric term g_{00} has opposite signs across the horizon $g_{00}(r = r^* - \epsilon) = -g_{00}(r = r^* + \epsilon) = \epsilon/r^*$. In such a situation, the constraint condition becomes $0 = E - \tilde{E}$, allowing its realization in the classical $\hbar \to 0$ limit. To a distant observer, the emitted radiation (made up of $r > r^*$ particles) is accompanied by an addition, to the black hole, of negative energy (due to the $r < r^*$ particles), that is, a loss of positive energy.

Such a thermal radiation has been shown by Hawking to have a blackbody temperature and a thermal energy inversely proportional to its mass M:

$$k_\mathrm{B}T = \frac{1}{8\pi}\frac{E_\mathrm{Pl}^2}{Mc^2}, \tag{A.10}$$

[3] A Killing vector ξ^μ characterizes a symmetry and its associated conservation law (along a geodesic). The conserved quantity in Eq. (6.43) is related to the Killing vector by $\partial L/\partial\dot{q} = -\xi^\mu g_{\mu\nu}(dx^\nu/d\tau)$.

where k_B is Boltzmann's constant, and $E_{Pl}^2 = \hbar c^5/G_N$ is the Planck energy squared (cf. (A.9)). Temperature's inverse proportionality with mass $T \sim M^{-1}$ means that for astrophysical black holes (large M) the Hawking radiation is rather weak and this emission is less significant than the in-falling material. On the other hand, for small-mass "mini black holes," because any loss of energy/mass will result in a hotter black hole (hence an even faster radiation), Hawking radiation will lead to an eventual total evaporation.

Entropy and black hole area increasing theorem Knowing the black-body temperature, one can associate a thermodynamic entropy S with a black hole (the **Bekenstein–Hawking entropy**) by a straightforward application of thermodynamic formulae, as $dS = T^{-1}dU$. Using (A.10) and $U = Mc^2$ for energy, we integrate the result to obtain the entropy associated with a black hole in units of Boltzmann's constant:

$$\frac{S}{k_B} = \frac{1}{4}4\pi\left(\frac{2G_N M}{c^2}\right)^2 \frac{c^3}{\hbar G_N} = \frac{1}{4}\frac{A^*}{l_{Pl}^2}, \tag{A.11}$$

where A^* is the horizon area $4\pi r^{*2} = 16\pi c^{-4}G_N^2 M^2$, and l_{Pl} is the Planck length. Since entropy is an ever-increasing function, the relation (A.11) between black hole entropy and area also implies that the black hole's horizon area is ever-increasing. This Bekenstein–Hawking "area increasing theorem" was in fact discovered before the advent of Hawking radiation. Even then it had been speculated that black hole formulae had a thermodynamic interpretation.

Black holes and advances in current study of quantum gravity Black holes are a unique arena to study quantum gravity. The singularity hidden behind the horizon demands both GR and quantum descriptions. Thus, a black hole is an ideal laboratory for thought experiments relating to quantum gravity. For instance, the Schwarzschild event horizon is seen to give rise to the phenomenon of infinite gravitational redshift, infinitely stretching the wavelength. Thus, Planck length phenomena near the horizon can be greatly amplified. It should be fruitful to investigate quantum gravitational processes associated with black holes. In this connection, we mention two important developments:

1. **The number of quantum states in a black hole.** Recall the well-known connection in statistical mechanics between entropy and information: $S = k_B \ln W$, where W is the number of microstates in the system under study. For a black hole, one can assume that W is given by the number of ways a quantum black hole can be formed. In string theory, for example, W has been calculated and found to be in perfect agreement with the Bekenstein–Hawking entropy of (A.11). That theories, such as superstring theory, give the correct count of quantum states of a black hole certainly encourages us to believe that they contain essentially correct ingredients for a true quantum gravity theory.
2. **The holographic principle.** The fact that the entropy of a black hole is proportional to its area, as in (A.11), has led to the conjecture that states in a spacetime region can equally well be represented by bits of information contained in its surface-boundary. This "holographic principle" by Gerard 't Hooft and Leonard Susskind has become one of the leading principles in the studies of theories of quantum gravity and unification.

A.3 False vacuum and hidden symmetry (Section 9.2.2)

In Section 9.2.2 we discussed the theoretical suggestion that the cosmological inflationary epoch is associated with a "false vacuum" of an inflation/Higgs field. This involves the concept of a "spontaneous breakdown of a symmetry," also described as a "hidden symmetry," for a symmetric theory having asymmetrical solutions. Namely, even though the theory is symmetric, its familiar symmetry properties are hidden. This can happen, as we shall see, when there are "degenerate ground states"—an infinite number of theoretically possible states (related to each other by symmetry transformations) all having the same lowest energy. But the physical vacuum is one of this set, and, by itself, it is not symmetric because it singles out a particular direction in the symmetry space. In this Appendix, we illustrate this phenomenon by the example of the breakdown of rotational symmetry in a ferromagnet near the Curie temperature.

A ferromagnet can be thought of as a collection of magnetic dipoles. When it is cooled below certain critical temperature, the Curie temperature T_c, it undergoes spontaneous magnetization: all its dipoles are aligned in one particular direction (a direction determined not by dipole interactions, but by external boundary conditions). Namely, when $T > T_c$ the ground state has zero magnetization $\vec{\mathcal{M}}_0 = 0$ because the dipoles are randomly oriented; but below the critical temperature $T < T_c$, all the dipoles line up, giving arise to a nonzero magnetization $\vec{\mathcal{M}}_0 \neq 0$, (Fig. A.2). This can happen even though the underlying dynamics of dipole–dipole interaction is rotationally symmetric—no preferred direction is built into the dynamics, that is, the theory has rotation symmetry.

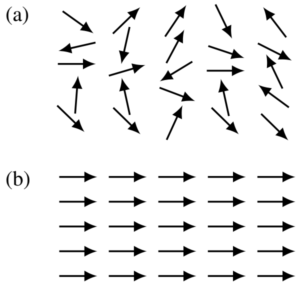

Fig. A.2 (a) Ground state with zero magnetization $\vec{\mathcal{M}}_0 = 0$ for randomly oriented dipoles. (b) Asymmetric ground state with $\vec{\mathcal{M}}_0 \neq 0$ because a particular direction is singled out.

For a mathematical description we shall follow the phenomenological theory of Ginzburg and Landau. When $T \approx T_c$, the rotationally symmetric free energy $\mathcal{F}(\vec{\mathcal{M}})$ of the system can be expanded in a power series of the magnetization $\vec{\mathcal{M}}$:

$$\mathcal{F}(\vec{\mathcal{M}}) = \left(\nabla_i \vec{\mathcal{M}}\right)^2 + \underbrace{a(T)(\vec{\mathcal{M}} \cdot \vec{\mathcal{M}}) + b(\vec{\mathcal{M}} \cdot \vec{\mathcal{M}})^2}_{V(\vec{\mathcal{M}})}. \tag{A.12}$$

In the potential energy function $V(\vec{\mathcal{M}})$ we have kept the higher order $(\vec{\mathcal{M}} \cdot \vec{\mathcal{M}})^2$ term, with a coefficient $b > 0$ (as required by the positivity of energy at large $\mathcal{M}$), because the coefficient a in front of the leading $(\vec{\mathcal{M}} \cdot \vec{\mathcal{M}})$ term can vanish: $a(T) = \gamma(T - T_c)$. With γ being some positive constant, the temperature-dependent coefficient a is positive when $T > T_c$, negative when $T < T_c$. Since the kinetic energy term $(\nabla_i \vec{\mathcal{M}})^2$ is non-negative, to obtain the ground state, we need only to minimize the potential energy:

$$\frac{dV}{d\vec{\mathcal{M}}} \propto \vec{\mathcal{M}} \left[a + 2b\left(\vec{\mathcal{M}} \cdot \vec{\mathcal{M}}\right)\right] = 0. \tag{A.13}$$

The solution of this equation gives us the ground state magnetization $\vec{\mathcal{M}}_0$. For $T > T_c$, hence a positive a, we get the usual solution of a zero magnetization $\vec{\mathcal{M}}_0 = 0$ (i.e. randomly oriented dipoles). This situation is shown in the plot of $V(\vec{\mathcal{M}})$ of Fig. A.3(a), where the potential energy surface is clearly symmetric with any rotation (in the 2D plane) around the central axis. (We have simplified the display to the case when $\vec{\mathcal{M}}$ is a 2D vector in a plane having two components $\mathcal{M}_1$ and $\mathcal{M}_2$.) However, for subcritical temperature $T < T_c$, the sign change of a brings about a change in the shape of the potential energy surface as in Fig. A.3(b). The surface remains symmetric with respect to rotation, but the zero magnetization point $\mathcal{M} = 0$ is now a local maximum. There are an infinite number of theoretically possible ground states at the bottom ring of the wine-bottle shaped surface—all having nonzero magnetization $\mathcal{M}_0 = \sqrt{-a/2b}$, but pointing in different directions in the 2D field space. These possible ground states are related to each other by rotations. The physical ground state, picked to be one of them by external conditions, singles out one specific direction, and hence is not rotationally symmetric. Below the Curie temperature, rotational symmetry in the ferromagnet is spontaneously broken and the usual symmetry properties of the underlying dynamics (in this case, rotational symmetry) are not apparent. We say spontaneous symmetry breaking corresponds to a situation of **hidden symmetry**.

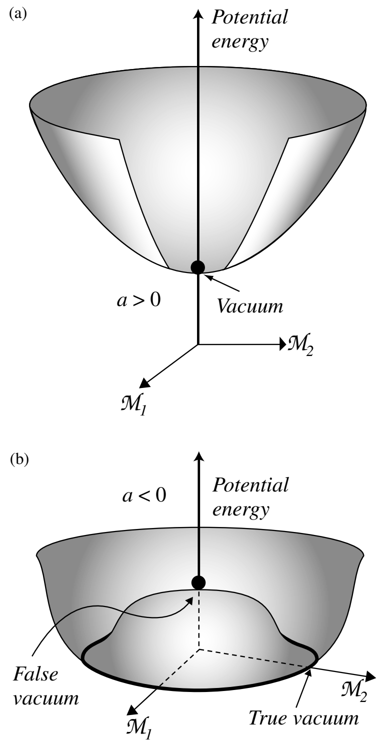

Fig. A.3 Symmetric potential energy surfaces in the 2D field space: (a) the normal solution, when the ground state is at a symmetric point with $\mathcal{M}_0 = 0$, and (b) the broken symmetry solution, when the energy surface has the shape of a "Mexican hat" with $\mathcal{M} = 0$ being a local maximum and the true ground state being one point in the trough (thus singling out one direction and breaking the rotational symmetry).

In particle physics we have a system of fields. Particularly it is postulated that there are scalar fields (for particles with zero spin) which have potential energy terms displaying the same spontaneous symmetry properties as ferromagnetism near T_c. The magnetization $\vec{\mathcal{M}}$ in (A.12) is replaced, in the case of particle physics, by a scalar field $\phi(x)$. Thus at high energy (i.e. high temperature) the system is in a symmetric phase (normal solution with $a(T) > 0$) and the unification of particle interactions is manifest (cf. the main text in this section); at lower energy (low temperature) the system enters a broken symmetry phase because of $a(T) < 0$. The ground state of a field system is, by definition, the vacuum. In this hidden symmetry phase we have a nonvanishing scalar field $\phi_0(x) \neq 0$. The relevance to cosmology is as follows: at higher temperature we have a symmetric vacuum. When the universe cools below the critical value, the same state becomes a local maximum and is at a higher energy than, and begins to roll toward, the true vacuum. We say the system (the universe) is temporarily, during the rollover period, in a false vacuum (cf. Fig. 9.3). This semiclassical description indicates the existence of constant field $\phi_0(x) \neq 0$ permeating everywhere in the universe.

A.4 The problem of quantum vacuum energy as Λ (Section 9.4)

It is very natural to identify the zero-point energy of the quantum fields as the cosmological constant Λ (the simplest form of the dark energy). Here we discuss very briefly the difficulty of such an association.

The quantum vacuum energy The introduction of the cosmological constant in the GR field equation does not explain its physical origin. In the inflation model it represents the false vacuum energy of an inflation/Higgs field. However, one natural contribution to Λ is the quantum mechanical vacuum energy[4] (also called the zero-point energy). From the view of quantum field theory, a vacuum state is not simply "nothingness." The uncertainty principle informs us that the vacuum has a constant energy density.[5]

The simplest way to see that a quantum vacuum state has energy is to start with the observation that the normal modes of a field are simply a set of harmonic oscillators. Summing over the quantized oscillator energies of all

[4] The inflationary cosmology discussion presupposes that the quantum vacuum contribution to the cosmological constant is negligibly small.

[5] In fact, QFT also pictures the vacuum as a sea of sizzling activities with constant creation and annihilation of particles.

the modes, we have

$$E_{\rm b} = \sum_i \left(\frac{1}{2} + n_i\right)\hbar\omega_i, \qquad \text{with } n_i = 0, 1, 2, 3, \ldots,. \tag{A.14}$$

(The subscript b stands for "boson." See later discussion.) From this we can identify the vacuum energy as

$$E_\Lambda = \sum_i \frac{1}{2}\hbar\omega_i. \tag{A.15}$$

At the atomic and subatomic levels, there is abundant empirical evidence for the reality of such a vacuum energy. For macroscopic physics, a notable manifestation of the zero point energy is the Casimir effect[6], which has been verified experimentally.

Natural size of quantum vacuum energy is 10^{120}-fold too large for Λ Nevertheless, a fundamental problem exists because the natural size of a quantum vacuum energy is enormous. Here is a simple estimate of the sum of (A.15). The energy of a particle with momentum k is $\sqrt{k^2c^2 + m^2c^4}$, see (10.36). From this we can calculate the sum by integrating over the momentum states to obtain the vacuum energy/mass density,

$$\rho_\Lambda c^2 = \int_0^\infty \frac{4\pi k^2 dk}{(2\pi\hbar)^3}\left(\frac{1}{2}\sqrt{k^2c^2 + m^2c^4}\right), \tag{A.16}$$

where $4\pi k^2 dk$ is the usual momentum phase space volume factor. This is a divergent quantity when we carry the integration to its infinity limit. Infinite momentum means zero distance; infinite momentum physics means zero distance scale physics. It seems natural that we should cut off the integral at the Planck momentum,

$$K_{\rm Pl} = \sqrt{\frac{\hbar c^3}{G_{\rm N}}} \simeq 10^{19}\,{\rm GeV}/c. \tag{A.17}$$

This is the scale when a quantum description of spacetime (i.e. quantum gravity) will be necessary (cf., Eq. (A.9)), and any GR singularities are expected to be modified. In this way, the integral (A.16) yields

$$\rho_\Lambda \cong \frac{K_{\rm Pl}^4}{16\pi^2\hbar^3 c} \simeq \frac{10^{74}\,{\rm GeV}^4}{c^2(\hbar c)^3} \simeq 2\times 10^{91}\ {\rm g/cm}^3 \tag{A.18}$$

This density is more than 10^{120} larger than the measured value of dark energy density, which is comparable to the critical density $\rho_\Lambda \approx \rho_{\rm c} = 2\times 10^{-29}$ g/cm^3.

Partial cancellation of boson and fermion vacuum energies We should note that in the above calculation, we have assumed that the field is a boson field (such as the photon and graviton fields), having integer spin and obeying Bose–Einstein statistics. The oscillator's creation and annihilation operators obey commutation relations, leading to symmetric wavefunctions. On the other hand, fermions (such as electrons, quarks, etc.) have half-integer spins, and obey Fermi–Dirac statistics. Their fields have normal modes behaving like **Fermi oscillators**. The corresponding creation and annihilation operators obey

[6] The summation of the modes in Eq. (A.15) involves the enumeration of the phase space volume in units of Planck's constant $\int d^3x d^3p (2\pi\hbar)^{-3}$, cf. Eq. (A.16). Since the zero-point energy has no dependence on position, one obtains a simple volume factor $\int d^3x = V$ and the result that the corresponding energy per unit volume $E_\Lambda V^{-1}$ is a constant with respect to changes in volume. As explained in Sec 9.1, this constant energy density implies a $-\partial E/\partial x$ force that is attractive, pulling-in the piston in Fig 9.1. This is the key property of the cosmological constant and is the origin of the Casimir effect — an attractive force between two parallel conducting plates.

anti-commutation relations, leading to antisymmetric wavefunctions. Such oscillators have a quantized energy spectrum as (see, for example, Das, 1993)

$$E_f = \sum_i \left(-\frac{1}{2} + n_i\right) \hbar\omega_i, \qquad \text{with } n_i = 0 \text{ or } 1 \text{ only.} \tag{A.19}$$

For a fermion field, the zero point energy is negative! Therefore, there will be a cancellation in the contributions by bosons and fermions. Many of the favored theories, attempting to extend the Standard Model of particle physics to the Planck scale, incorporate the idea of **supersymmetry**. In such theories the bosonic and fermionic degrees of freedom are equal. In fact, the vacuum energy of systems with exact supersymmetry must vanish (i.e. an exact cancellation). However, we know that in reality supersymmetry cannot be exact because its implication of equal boson and fermion masses $m_F = m_B$ in any supersymmetric multiplet is not observed in nature. (For example, we do not see a spin-zero particle, a "selectron," having the same properties, and degenerate in mass, as the electron; similarly we have not detected the photon's superpartner, a massless spin-$\frac{1}{2}$ particle called "photino," etc. A plausible interpretation is that they are much more massive and are yet to be produced and detected in our high energy accelerators.) So this supersymmetry must be broken and we expect only a partial cancellation between bosons and fermions. The cosmological constant problem, in this context, is the puzzle why such a fantastic cancellation takes place: a cancellation of the first 120 significant figures (yet stops at the 121st place)! If it merely reflects a broken supersymmetry, one would still have a vacuum energy, by comparing the first-order fermion and boson contributions in Eq. (A.16), leading to a result that modifies the boson Eq. (A.18) as

$$\rho_\Lambda \cong \frac{K_{\mathrm{Pl}}^4}{16\pi^2\hbar^3 c}\left(\frac{\Delta m^2 c^2}{K_{\mathrm{Pl}}^2}\right), \tag{A.20}$$

where Δm^2 is the fermion and boson mass difference $m_F^2 - m_B^2$. But the phenomenologically allowed value of $\Delta m^2 \gtrsim (10^2 \text{ GeV}/c)^2$ can only produce a suppression factor of $(\Delta m^2 c^2 / K_{\mathrm{Pl}}^2) = 10^{-36}$ at the most—thus still some 80 to 90 orders short of the required $O(10^{-120})$. Clearly, something fundamental is missing in our understanding of the physics behind the cosmological constant.

The energy scale associated with a quantum description of the dark energy Another simple way of looking at this cosmological constant problem is as follows: Assuming that the above-discussed quantum zero-point energy is somehow absent, Eq. (A.18) is then interpreted as a dimensional analysis of the cosmological constant if a quantum description is involved—$E_{\mathrm{Pl}} = cK_{\mathrm{Pl}}$ being then taken as the energy scale E_X of whatever physics is associated with the dark energy. The observed dark energy, being comparable to the critical density $\rho_c c^2$ (cf. Eq. (7.19)), then corresponds to an energy scale of

$$E_X \cong [16\pi^2\hbar^3 c^5 \rho_c]^{1/4} = O(10^{-3} \text{ eV}). \tag{A.21}$$

Phrased in this way the cosmological constant problem is, there is no known physics that one can naturally associate it with such an energy scale.

Answer keys to review questions

B

Here we provide sketchy answers to the review questions presented at the end of each chapter.

Chapter 1: Introduction and overview

1. Relativity is the coordinate symmetry. SR is the one with respect to inertial frames; GR, to general coordinate frames.
2. Covariance of the physics equations (i.e. physics is not changed) under symmetry transformations. Not able to detect a particular physical feature means physics is unchanged. It is a symmetry. If one cannot detect the effect after changing the orientation it means the physics equation is covariant under rotation—thus rotationally symmetric. Similar statements can be made for the coordinate symmetry of relativity.
3. A tensor is a mathematical object having definite transformation properties under coordinate transformation. A tensor equation, having the same transformation for every term, maintains the same form and is thus symmetric under coordinate changes.
4. (a) The frames in which Newton's first law holds. (b) The frames moving with constant velocity with respect to the fixed stars. (c) The frames in which gravity is absent.
5. Equations of Newtonian physics are covariant under Galilean transformations; electrodynamics, under Lorentz transformations. Galilean transformations are the Lorentz transformations with low relative velocity.
6. The coordinate transformations in GR are necessarily position-dependent.
7. See Section 1.2.1.
8. In GR the gravitational field is the curved spacetime. The Einstein equation is the GR field equation. The geodesic equation is the GR equation of motion.
9. In Newtonian physics, space is the arena in which physical events take place. In GR, space simply reflects the relationship among physical events taking place in the world and has no independent existence.

Chapter 2: Special relativity and the flat spacetime

1. Equations (2.13) and (2.10).
2. First paragraph in Box 2.2.

3. $\Delta s^2 = \Delta \mathbf{x}^2 - c^2 \Delta t^2$ is invariant and "absolute"—the same value to all observers, because Δs is the interval of proper time. There is only one rest frame; all observers must agree on this value. Δs is the "length" because $\Delta s^2 = g_{\mu\nu} \Delta x^\mu \Delta x^\nu$ with $\Delta x^0 = c\Delta t$. And a length invariant transformation is a rotation.
4. Einstein clarified the meaning of time measurement when the signal transmission could not be instantaneous and showed that $\Delta t' \neq \Delta t$ was physical.
5. (a) $\mathbf{e}_i \cdot \mathbf{e}_j \equiv g_{ij}$, and (b) $ds^2 = g_{ij} \Delta x^i \Delta x^j$. Diagonal elements are lengths of basis vectors and off-diagonal elements the deviation from orthogonality. Cartesian system: $g_{ij} = \delta_{ij}$.
6. $[\bar{\mathbf{R}}^{\mathsf{T}}][\mathbf{g}][\bar{\mathbf{R}}] = [\mathbf{g}]$ reduces to $[\bar{\mathbf{R}}^{\mathsf{T}}][\bar{\mathbf{R}}] = [\mathbf{1}]$ in the Cartesian coordinate with $[\mathbf{g}] = [\mathbf{1}]$.
7. Cf. (2.60).
8. Figure 2.6.
9. "simultaneity is relative" means that it is possible to have $\Delta t \neq 0$ even though $\Delta t' = 0$. See Fig. 2.14.
10. See Fig. 2.9.
11. $t_1 = t_2$, thus $\Delta t = 0$ in (2.65).
12. $x_1' = x_2'$, thus $\Delta x' = 0$ in (2.65).
13. See Fig. 2.9.

Chapter 3: The principle of equivalence

1. Newton's field equation is (3.6). The equation of motion is (3.8), which is totally independent of any properties of the test particle.
2. Equations (3.9) and (3.10). Experimental evidence: $(\ddot{\mathbf{r}})_{\mathrm{A}} = (\ddot{\mathbf{r}})_{\mathrm{B}}$ for any two objects A and B.
3. The equivalence principle (EP) is the statement that the physics in a freely falling frame in a gravitational field is indistinguishable from the physics in an inertial frame without gravity. Weak EP is EP restricted only to Newtonian mechanics; strong EP is EP applied to all physics including electrodynamics.
4. Cf. Fig. 3.3.
5. (a) Equation (3.22); (b) Eq. (3.52).
6. (a) Equation. (3.33); (b) Eq. (3.36).
7. Cf. Eq. (3.38).
8. Cf. Eq. (3.40). The speed of light is absolute as long as it is measured with respect to the proper time of the observer. It may vary if measured by the coordinate time. Example: light deflection in a gravitational field.

Chapter 4: Metric description of a curved space

1. An "intrinsic geometric description" is one that an inhabitant living within the space can perform without referring to any embedding.
2. The metric elements can be found by distance measurement once the coordinates have been fixed. See Eq. (4.16).
3. The geodesic equation is the Euler–Lagrange equation resulting from extremization of the length integral $\int g_{ab}\, \dot{x}^a\, \dot{x}^b\, d\sigma$.

4. The metric is an intrinsic quantity because it can be determined through intrinsic operations (cf. Question 2). Other intrinsic geometric quantities, such as angle, geodesic curves, etc. can then be derived from the metric function.
5. A flat surface can also have a position-dependent metric. Example: polar coordinates on a flat plane.
6. Transformation in a curved space must be position-dependent.
7. A small region of any curved space can be approximated by a flat space: can always find a coordinate transformation at a given point so that the new metric is a constant up to second-order correction.
8. K vanishes only for a flat surface independent of coordinate choices, and it measures the deviation from Euclidean geometric relations.
9. 2-sphere, 2-pseudosphere, and flat plane.
10. In a 4D Euclidean space $W^2 + X^2 + Y^2 + Z^2 = R^2$ describes a 3-sphere; in a 4D Minkowski space $-W^2 + X^2 + Y^2 + Z^2 = -R^2$, a 3-pseudosphere.
11. The circumference of a circle is related to its radius r as $S = 2\pi r + \pi r^3 K/3$ for small r.
12. Cf. (4.50) and Fig. 4.7(a).

Chapter 5: GR as a geometric theory of gravity - I

1. In a "geometric theory of physics" the physical results can be directly attributed to the underlying geometry of space and time. Example: latitudinal distances decrease as they approach either of the poles, cf. Fig. 4.2, reflects the geometry of the spherical surface (rather than the physics of ruler lengths). See Section 5.1.
2. Time interval changes because the spacetime metric is position dependent, cf. (5.2) with the metric element being directly related to the gravitational potential, $g_{00} = -(1 + 2\Phi/c^2)$.
3. The relation between circumference and radius deviates from the Euclidean $S = 2\pi R$, cf. Section 5.1.1.
4. Curved spacetime is the gravitational field. See Section 5.1.2.
5. Space and time are described by a metric function, which is the solution to the GR field equation.
6. We expect the particle trajectory to be the straightest and shortest possible worldline in spacetime.
7. Newtonian limit: $v \ll c$ particles in a weak and static gravitational field. Cf. Eqs (5.23) and (5.25).
8. The tidal forces are the relative gravitational forces on neighboring test particles. They are the second derivatives of gravitational potential, thus $\propto r^{-3}$ for a spherical source. The extra power in the denominator means that solar tidal force can be smaller than lunar ones even though the leading gravitational force due to the sun is larger. Relativistic gravitational potential being the metric, the tidal forces must be second derivatives of the metric, hence the curvature.
9. GR field equation has the structure of the curvature being proportional to the energy–momentum tensor $G_{\mu\nu} = \kappa T_{\mu\nu}$ with the proportionality constant $\kappa \sim G_{\mathrm{N}}$. The curvature and energy density having different

measurement units, Newton's is the basic "conversion factor" in GR. Light speed c in SR (space–time symmetry), and Planck's constant h in quantum theory (wave–particle duality).

10. Box 5.3

Chapter 6: Spacetime outside a spherical star

1. See Eq. (6.12). Curved space because $d\rho = \sqrt{g_{rr}}\,dr$ and time because $d\tau = \sqrt{-g_{00}}\,dt$.
2. Newtonian gravity for a spherical source is the same as if all the mass is concentrated at the center. Thus, time dependence of source does not show up in the exterior field, and there is no monopole radiation.
3. $g_{00} = -1 + r^*/r = -(1 + 2\Phi/c^2)$. Thus $\Phi = -G_\mathrm{N}M/r$.
4. Cf. Eqs (6.26) and (6.27).
5. When the source, lensing mass, and observer are perfectly aligned, the resultant azimuthal symmetry leads to an "Einstein ring." When the lensing mass is not huge, the separations among multiple images are small, and the images cannot be resolved. This results in the overlap of images and an enhancement of the brightness.
6. Equation (6.50).
7. It disappears in some coordinates such as the Eddington–Finkelstein system.
8. (a) Infinite gravitational time dilation (6.77), and (b) lightcones tipping over, forcing all worldlines toward $r = 0$.
9. Through their gravitational effects on other objects. Evidence for black holes in (a) X-ray binaries, (b) supermassive galactic centers.

Chapter 7: The homogeneous and isotropic universe

1. $v = H_0 r$ is linear if H_0 is independent of v and r. This means that every galaxy sees all other galaxies as rushing away according to Hubble's law (cf. discussion relating to Fig. 7.3).
2. $t_\mathrm{H} = H_0^{-1} = t_0$ for an empty universe. In a universe full of matter and energy we expect $t_\mathrm{H} > t_0$ because the expansion rate was faster in the past as gravitational attraction has been slowing down the expansion. $[t_0]_\mathrm{gc} \gtrsim 12.5$ Gyr.
3. The "rotation curves" are the plots of matter's rotational speed as a function of radial distance to the center of the mass distribution (e.g. a galaxy). Gravitational theory would lead us to expect a rotational curve to drop as $v \sim r^{-1/2}$ outside the matter distribution. Rotation curves are observed to stay flat, $v \sim r^0$, way beyond the luminous matter distribution.
4. A simple example: two masses $M \gg m$ illustrates the content of the virial theorem $\langle V \rangle = -2\langle T \rangle$ for a gravitational system: $G_\mathrm{N}M\langle s^{-1}\rangle = \langle v^2 \rangle$.
5. $\Omega_\mathrm{M} \simeq 0.3$, $\Omega_\mathrm{LM} \lesssim 0.005$, and $\Omega_\mathrm{B} \simeq 0.04$. Thus $\Omega_\mathrm{exotic} = \Omega_\mathrm{M} - \Omega_\mathrm{B} \simeq 0.26$.

6. The cosmological principle: at any given instance of cosmic time, the universe appears the same at every point: space is homogeneous and isotropic. The comoving coordinates are a system where the time coordinate is chosen to be the proper time of each cosmic fluid element; the spatial coordinates are coordinate labels carried along by each fluid element (thus each fluid element has a fixed and unchanging comoving spatial coordinate).
7. See Eqs (7.38), (7.42), and (7.43) for the Robertson–Walker metric in the spherical polar and cylindrical coordinate systems. The input used in the derivation is the cosmological principle.
8. $a(t)$ is the dimensionless scale which changes along with the cosmic time, and $k = \pm 1, 0$ is the curvature parameter. The Hubble constant is related to the scale factor by $H(t) = \dot{a}(t)/a(t)$.
9. The scaling behavior of wavelength being
$$\frac{\lambda_{\rm rec}}{\lambda_{\rm em}} = \frac{a(t_{\rm rec})}{a(t_{\rm em})},$$
the definition of redshift $z \equiv (\lambda_{\rm rec} - \lambda_{\rm em})/\lambda_{\rm em}$ leads to
$$\frac{a(t_0)}{a(t_{\rm em})} = 1 + z.$$
10. Summarize the derivation of (7.49), (7.51), and (7.55): starting from the $d\Omega = 0$ invariant separation $ds^2 = -c^2\,dt^2 + R_0^2 a^2(t)\,dr^2$, the proper distance for $dt = 0$ can be obtained by straightforward integration over ds. The result has the form of $d_{\rm p}(r,t) = R_0 a(t) r$. The proper distance measured at t_0 to a light source at r is the light ($ds^2 = 0$) path given by $d_{\rm p}(r, t_0) = R_0 r = \int_{t_{\rm em}}^{t_0} c dt/a(t)$. The time integral can be converted to one over redshift:
$$\int_{t_{\rm em}}^{t_0} \frac{cdt}{a(t)} = \int_a^1 \frac{cda}{a\dot{a}} = \int_a^1 \frac{cda}{a^2 H} = \int_0^z \frac{cdz}{H}.$$
11. Luminosity distance is defined through the observed flux in relation to the intrinsic luminosity of the source $d_{\rm L} = \sqrt{L/4\pi f}$, and is related to proper distance by $d_{\rm L} = (1 + z) d_{\rm p}$.

Chapter 8: The expanding universe and thermal relics

1. The Friedmann equation is the Einstein equation subject to the cosmological principle, that is, the Robertson–Walker metric and ideal fluid $T_{\mu\nu}$. Newtonian interpretation: energy balance equation $E_{\rm tot}$ being the sum of kinetic and potential energies. Newtonian interpretation is possible because of cosmological principle—large region behaves similarly as the small. Only quasi-Newtonian because we still need to supplement it with geometric concepts like curvature and scale factor, etc.
2. Both the critical density and escape velocity are used to compare the kinetic and potential terms to determine whether the total energy is positive (bound system) or negative (unbound system).

3. Radiation energy, being proportional to frequency (hence inverse wavelength), scales as a^{-1}. After dividing by the volume (a^3), the density $\sim a^{-4}$. Since matter density scales as a^{-3}, radiation dominates in the early universe.
4. $p = w\rho c^2$ with $w_\mathrm{R} = 1/3$ and $w_\mathrm{M} = 0$. Flat RDU $a \sim t^{1/2}$ and $t_0 = \frac{1}{2}t_\mathrm{H}$; MDU $a \sim t^{2/3}$ and $t_0 = \frac{2}{3}t_\mathrm{H}$. Since radiation–matter equality time $t_\mathrm{RM} \ll t_{0,}$ MDU should be a good approximation.
5. Figure 8.2.
6. Stefan–Boltzmann law: $\rho_\mathrm{R} \sim T^4$ and radiation density scaling law $\rho_\mathrm{R} \sim a^{-4}$. Therefore $T \backsim a^{-1}$. Blackbody radiation involves only scale invariant combinations of (volume)$\times E^2 dE$ (recall radiation energy $\backsim a^{-1}$) and E/T.
7. Reaction rate faster than expansion rate. Cf. (8.45) and (8.46).
8. From $E_\mathrm{bbn} \approx$ MeV we have $T_\mathrm{bbn} \approx 10^9\,\mathrm{K}$. Boltzmann distribution yields $n_\mathrm{n}/n_\mathrm{p} \simeq \exp[-\Delta mc^2/k_\mathrm{b}T_\mathrm{bbn}]$ with $\Delta m = m_\mathrm{n} - m_\mathrm{p}$. Equation (8.55) leads to mass fraction of $\frac{1}{4}$, if $n_\mathrm{n}/n_\mathrm{p} \backsimeq \frac{1}{7}$.
9. Because the theoretical prediction of deuterium abundance by big bang nucleosynthesis is sensitive to Ω_B and number of neutrino flavors. The observed abundance can fix these quantities, cf. Fig. 8.3.
10. At t_γ the reversible reaction of $\mathrm{e} + \mathrm{p} \longleftrightarrow \mathrm{H} + \gamma$ stopped proceeding from right to left. All charged particles turned into neutral atoms. The universe became transparent to photons. Average thermal energy $O\,(\mathrm{eV})$ translates into $T(z_\gamma) \simeq 3{,}000\,\mathrm{K}$. By the temperature scaling law
$$\frac{T(t_\gamma)}{T(t_0)} = \frac{a(t_0)}{a(t_\gamma)} = 1 + z_\gamma$$
leading to $T(t_0) \simeq 3\,\mathrm{K}$.
11. $t_\mathrm{bbn} \simeq 10^2\,\mathrm{s}$, $t_\mathrm{RM} \simeq 10{,}000$ year, and $t_\gamma \simeq 350{,}000$ year.
12. See discussion leading to (8.77).
13. Motion leads to frequency blueshift in one direction and redshift in the opposite direction. Frequency shift means energy change, hence temperature change.
14. Primordial density perturbation as amplified by gravity and resisted by radiation pressure set up acoustic waves in the photon–baryon fluid. Photons leaving denser regions would be gravitationally redshifted and thus bring about CMB temperature anisotropy.
15. Because cosmological theories predict only statistical distribution of hot and cold spots. To compare theory to experiment we need to average over an ensemble of identically prepared universes. Having only one universe, all we can do is to average samples from regions corresponding to different m moment number (for the same l). But for a given l, there are only $2l + 1$ values of m. For low l distributions there is large variance. Cf. (8.89).

Chapter 9: Inflation and the accelerating universe

1. Constant density means $dE = \varepsilon dV$. The first law $dE = -pdV$ leads to $p = -\varepsilon$.

2. Cf. Section 9.1.1.
3. Cf. Eq. (9.16).
4. We see the same CMB temperature across patches of the sky $\gg 1°$ which is the horizon angular separation at t_γ. Since they could not have been causally connected and thermalized, how could they have the same property?
5. Cf. Fig. 9.3.
6. Repulsive expansion by the constant energy density is self-reinforcing: the more the volume increases the more gain in energy, leading in turn to more repulsive expansion, $\dot{a}(t) \propto a(t)$. Rapid expansion stretches out any curvature, solving the flatness problem. Also because it is possible to have $\dot{a}(t) > c$, one thermalized volume could be stretched out into such a large region with many horizon lengths across, resolving the horizon puzzle.
7. All came from the potential energy of inflation/Higgs field which turned into the false vacuum energy during the phase transition. Quantum fluctuation of the Higgs field became the density fluctuations that seeded the cosmic structure. Cf. Section 9.2.3.
8. (I) The large angle region ($>1°$): we see the initial density perturbation. (II) The subdegree region: we see the signals of the acoustic waves of photon–baryon fluid, with gravity the driving force and radiation pressure the restoring force. (III) The small angle (less than arc-minute) region: photon decoupling shows up as exponential damping during this finite interval of small intervals.
9. The primary peak should correspond to the fundamental wave with a wavelength given by the sound horizon of the photon-baryon fluid. The corresponding angular separation is the sound horizon length at t_γ divided by the angular diameter distance from t_γ to us at t_0. This diameter distance would be affected by the curvature of space. The observation that the first peak at $l \approx 200$ agrees with the prediction of a flat universe.
10. An accelerating expansion means that expansion was slower in the past, hence a longer age of the universe.
11. Cf. sidebar 5, p. 184.
12. Their intrinsic luminosities can be reliably calibrated, and they are extremely bright.
13. An accelerating expansion means that expansion was slower in the past. It would take a longer period, thus longer separation, before reaching a given redshift (recession velocity). The longer distance to the galaxy translates into a dimmer light.
14. If we live in an accelerating universe powered by the constant energy density of the cosmological constant, at earlier epochs the universe must be dominated by ordinary radiation and matter: instead of being a constant, ρ_R and ρ_M increases as a^{-4} and a^{-3} as $a \to 0$. Thus the accelerating phase must be preceded by deceleration.
15. Why should $\Omega_M \simeq \Omega_\Lambda$ now, or equivalently the matter-Λ equality time $t_{M\Lambda} \simeq t_0$?
16. See bullet summary at the beginning of the chapter.

Chapter 10: Tensors in special relativity

1. The covariant components are the projection of the vector onto the basis vectors $V_\mu = \mathbf{V} \cdot \mathbf{e}_\mu$ and the contravariant components onto the inverse bases $V^\mu = \mathbf{V} \cdot \mathbf{e}^\mu$. These two kinds of vector (tensor) components are needed to construct invariants such as $V_\mu U^\mu$.
2. $$\begin{pmatrix} ct' \\ x' \end{pmatrix} = \begin{pmatrix} \gamma & -\beta\gamma \\ -\beta\gamma & \gamma \end{pmatrix} \begin{pmatrix} ct \\ x \end{pmatrix}$$ $$\begin{pmatrix} c^{-1}\partial'_t \\ \partial'_x \end{pmatrix} = \begin{pmatrix} \gamma & \beta\gamma \\ \beta\gamma & \gamma \end{pmatrix} \begin{pmatrix} c^{-1}\partial_t \\ \partial_x \end{pmatrix}$$
3. The position 4-vector x^μ is naturally contravariant, as opposed to covariant with components $x_\mu = (-ct, \mathbf{x})$ and the del operator ∂_μ naturally covariant while the contravariant del components $\partial^\mu = (-c^{-1}\partial_t, \nabla)$ are "unnatural."
4. Contravariant and covariant vectors transform differently. $$V'^\mu = [\mathbf{L}]^\mu_{\ \nu} V^\nu,$$ $$V'_\mu = [\bar{\mathbf{L}}]_\mu^{\ \nu} V_\nu,$$ where $[\bar{\mathbf{L}}] = [\mathbf{L}]^{-1}$.
5. $$T'^{\mu}_{\ \nu} = [\mathbf{L}]^\mu_{\ \lambda} [\bar{\mathbf{L}}]_\nu^{\ \rho} T^\lambda_{\ \rho}$$
6. dx^μ/dt is not a 4-vector because t is not a scalar. The velocity 4-vector $dx^\mu/d\tau = \gamma(dx^\mu/dt)$.
7. The relativistic energy $E = \gamma mc^2$ and 3-momentum $\mathbf{p} = \gamma m\mathbf{v}$. With $p^\mu = m(dx^\mu/d\tau) = ((E/c), \mathbf{p})$, we have the invariant $p^\mu p_\mu = -m^2c^2 = -(E/c)^2 + \mathbf{p}^2$.
8. Given $\partial_\mu F^{\mu\nu} + j^\nu/c = 0$ we can show $\partial'_\mu F'^{\mu\nu} + j'^\nu/c = 0$ $$\left(\partial' F' + j'/c\right) = [\bar{\mathbf{L}}]\partial[\mathbf{L}][\mathbf{L}]F + [\mathbf{L}]\,j/c = [\mathbf{L}](\partial F + j/c) = 0$$ Namely, because every term is a 4-vector, the form of the equation is unchanged under the Lorentz transformation. This equation includes the statement of electric charge conservation $\partial_\mu j^\mu = 0$ because $\partial_\mu \partial_\nu F^{\mu\nu} = 0$ as $F^{\mu\nu} = -F^{\nu\mu}$ and $\partial_\mu \partial_\nu = \partial_\nu \partial_\mu$.
9. T^{00} = energy density, T^{0i} = momentum density or energy current-density, and T^{ij} = normal force per unit area (pressure) for $i = j$, shear force for unit area for $i \neq j$.
10. Equation (10.88).

Chapter 11: Tensors in general relativity

1. Equations (11.8) and (11.13).
2. Transformations in a curved space must necessarily be position-dependent.
3. Equations (11.16) and (11.19). Tensor equations are automatically relativistic.

4. $V^\mu = \mathbf{e}^\mu \cdot \mathbf{V}$, being coordinate dependent, their derivative will have extra term $\sim (\partial_\nu \mathbf{e}^\mu)$ as in (11.18).
5. $D_\nu T^{\lambda\rho}_\mu = \partial_\nu T^{\lambda\rho}_\mu - \Gamma^\sigma_{\nu\mu} T^{\lambda\rho}_\sigma + \Gamma^\lambda_{\nu\sigma} T^{\sigma\rho}_\mu + \Gamma^\rho_{\nu\sigma} T^{\lambda\sigma}_\mu$.
6. Equation (11.37).
7. We can express the Riemann tensor as a commutator of covariant derivatives (11.67). Since every other term is known to be a good tensor, by the quotient theorem $R^\mu{}_{\lambda\alpha\beta}$ must also be a good tensor.
8. $G^{\mu\nu} = G^{\nu\mu}$ and $D_\mu G^{\mu\nu} = 0$.
9. At every point one can always find a coordinate system (LEF) in which $\partial_\mu g_{\nu\lambda} = 0$ and $\Gamma^\lambda_{\nu\sigma} = 0$. Therefore, we have $D_\mu g_{\nu\lambda} = 0$ in the LEF. Since this is a tensor equation, it must be valid in every frame.

Chapter 12: GR as a geometric theory of gravity - II

1. See first part of Section 12.1.
2. Just replace ordinary derivatives by covariant derivatives. This is required because the coordinate symmetry in GR is local. It involves position-dependent transformations. Covariant derivatives include the Christoffel symbols which, being the derivatives of the gravitational potential (i.e. the metric), constitute the gravitational field.
3. Following the procedure stated in Question 2, EOM in SR $(dU^\mu/d\tau) = 0$ leads to $(DU^\mu/D\tau) = (dU^\mu/d\tau) + \Gamma^\mu_{\nu\lambda} U^\nu (dx^\lambda/d\tau) = 0$.
4. Section 12.2.1.
5. Equations (12.26) and (12.27).
6. Set the Ricci scalar for the 3D spatial metric for a spherically symmetric space to a constant, because homogeneous and isotropic space corresponds to a space having spherical symmetry with respect to every point. Namely, a homogeneous and isotropic space must be a space with constant curvature.
7. Friedmann equations are components of just the Einstein equation with Robertson–Walker metric and an energy-momentum tensor given by that of an ideal fluid.
8. $g^{\mu\nu} = g^{\nu\mu}$ and $D_\mu g^{\mu\nu} = 0$, cf. Question 8 in Chapter 11.
9.
$$G_{\mu\nu} = \kappa(T_{\mu\nu} + \kappa^{-1}\Lambda g_{\mu\nu}) = \kappa(T_{\mu\nu} + T^\Lambda_{\mu\nu}).$$

   Thus, even in the absence of matter/energy source $T_{\mu\nu} = 0$, space can still be curved by the Λ term.

Chapter 13: Linearized theory and gravitational waves

1. Newton's is a static theory: it has no nontrivial time dependence. Einstein's theory, being relativistic, treats space and time on an equal footing, hence has nontrivial time dependence.

2. This new window to the universe allows us to observe strong gravity regions which are often at the core of many interesting astrophysical phenomena.
3. Metric is slightly different from flat Minkowski metric $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$ with $h_{\mu\nu} \ll 1$. All GR equations taken only to the first order in $h_{\mu\nu}$. Propagation of a gravitational wave can be viewed as ripples of curvature moving in the Minkowski spacetime.
4. Both are long range forces. Their quanta are massless. EM waves (photon) have spin 1 and gravitational waves (graviton) spin 2. Leading EM radiation is dipole, gravitational radiation is quadrupole.
5. Coordinate transformations among the slightly curved coordinates. $\partial^{\mu}\bar{h}_{\mu\nu} = 0$ is the Lorentz gauge condition. Can make further gauge transformations as long as the vector gauge function satisfies the $\Box\chi_{\mu} = 0$ constraint.
6. Figure 13.1.
7. Equation (13.41)
8. Poynting vector $\mathbf{S} = \mathbf{E} \times \mathbf{B} \propto (\dot{\mathbf{A}})^2$ where $\mathbf{A}$ is the EM vector potential. We expect the gravitational wave energy flux to be proportional to $\dot{h}^2$ also. The dimensionful proportionality constant can be fixed by dimensional analysis to be c^3/G_{N}.
9. No monopole by Birkhoff's theorem. The amplitude must be the second derivative of the mass moments. No dipole radiation because the dipole amplitude is just the rate of total momentum change, which vanishes for an isolated system, cf. Eq. (13.60).
10. See the first sidenote in this chapter for a description of pulsars. From the observed increase in the rate of pulse arrival time, one can deduce that the binary orbit period is decreasing, which matches perfectly the GR prediction of energy loss due to gravitational wave emission by the circulating system.

Solutions of selected problems

C

(2.2) **Inverse Lorentz transformation** The Lorentz transformation (2.73), and its inverse, written out only for the nontrivial components, are

$$\begin{pmatrix} ct' \\ x' \end{pmatrix} = \gamma \begin{pmatrix} 1 & -\beta \\ -\beta & 1 \end{pmatrix} \begin{pmatrix} ct \\ x \end{pmatrix},$$
$$\begin{pmatrix} ct \\ x \end{pmatrix} = \gamma \begin{pmatrix} 1 & \beta \\ \beta & 1 \end{pmatrix} \begin{pmatrix} ct' \\ x' \end{pmatrix}. \tag{C.1}$$

The inverse matrix relation is demonstrated by

$$\gamma^2 \begin{pmatrix} 1 & -\beta \\ -\beta & 1 \end{pmatrix} \begin{pmatrix} 1 & \beta \\ \beta & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

after using $\gamma^2 = (1 - \beta^2)^{-1}$.

(2.3) **Lorentz transformation of derivative operators**

(a) Starting with the chain rule,

$$\frac{\partial}{\partial x'} = \frac{\partial x}{\partial x'}\frac{\partial}{\partial x} + \frac{\partial t}{\partial x'}\frac{\partial}{\partial t} = \gamma \frac{\partial}{\partial x} + \gamma\beta \frac{\partial}{c\partial t}.$$

To reach the last equality, we have used (C.1) showing (ct, x) as functions of (ct', x') to calculate $\partial x/\partial x'$ and $\partial t/\partial x'$. Similarly, we have

$$\frac{\partial}{c\partial t'} = \gamma \frac{\partial}{c\partial t} + \gamma\beta \frac{\partial}{\partial x}.$$

(b) $[\bar{\mathbf{L}}]$ can be found by substituting into $\delta^\nu_\mu = \partial(x'^\nu)/\partial x'^\mu \equiv \partial'_\mu x'^\nu$ the respective Lorentz transformations $[\mathbf{L}]$ and $[\bar{\mathbf{L}}]$ for coordinates and coordinate derivatives Eqs (2.74) and (2.75):

$$\delta^\nu_\mu = \partial'_\mu x'^\nu = \sum_{\lambda,\rho} ([\bar{\mathbf{L}}]^\lambda_\mu \partial_\lambda)([\mathbf{L}]^\nu_\rho x^\rho) = \sum_\lambda [\bar{\mathbf{L}}]^\lambda_\mu [\mathbf{L}]^\nu_\lambda. \tag{C.2}$$

Namely, $\mathbf{1} = [\bar{\mathbf{L}}][\mathbf{L}]$. Thus, the transformation for the derivative is just the inverse shown in (C.1).

(2.4) **Lorentz covariance of Maxwell's equations** Given (2.78), we show the validity of (2.77) by applying the Lorentz transformations for the fields and spacetime derivatives:

$$\begin{aligned}\nabla'\cdot\mathbf{B}' &= \frac{\partial B'_x}{\partial x'} + \frac{\partial B'_y}{\partial y'} + \frac{\partial B'_z}{\partial z'}\\ &= \gamma\left(\frac{\partial}{\partial x} + \beta\frac{\partial}{c\partial t}\right)B_x + \frac{\partial}{\partial y}\gamma(B_y + \beta E_z) + \frac{\partial}{\partial z}\gamma(B_z - \beta E_y)\\ &= \gamma\underbrace{\left(\frac{\partial B_x}{\partial x} + \frac{\partial B_y}{\partial y} + \frac{\partial B_z}{\partial z}\right)}_{\nabla\cdot B=0} + \gamma\beta\underbrace{\left[\frac{\partial B_x}{c\partial t} + \left(\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z}\right)\right]}_{((1/c)(\partial\mathbf{B}/\partial t)+\nabla\times\mathbf{E})_x=0},\end{aligned} \tag{C.3}$$

where we have used the Lorentz transformation of (2.76) and (2.18) to reach the second line. The x-component of Faraday's equation being singled out is due to the fact that we have assumed a Lorentz boost in the x direction.

(2.5) **From Coulomb's to Ampere's law** To derive Faraday's law from the magnetic Gauss's we note that given the validity of the magnetic Gauss laws in both frames $\nabla\cdot\mathbf{B} = 0$ and $\nabla'\cdot\mathbf{B}' = 0$, (C.3) then implies that the x component of the equation $\nabla\times\mathbf{E} + (1/c)(\partial\mathbf{B}/\partial t) = 0$. Hence the vector equation $\nabla\times\mathbf{E} + (1/c)(\partial\mathbf{B}/\partial t) = 0$ as the y and z components can be similarly deduced by considering Lorentz boosts in y and z directions.

(2.8) **Group property of Lorentz transformations** Only display the group property of the boost transformation: given the Lorentz boost (2.57), we have the combined transformation

$$[\mathbf{L}(\psi_1)][\mathbf{L}(\psi_2)] = \begin{pmatrix} c_1 & s_1\\ s_1 & c_1\end{pmatrix}\begin{pmatrix} c_2 & s_2\\ s_2 & c_2\end{pmatrix},$$

where $c_1 \equiv \cosh\psi_1$ and $s_1 \equiv \sinh\psi_1$. A straightforward matrix multiplication and the trigonometric identities, with $c_{12} \equiv \cosh(\psi_1 + \psi_2)$ and $s_{12} \equiv \sinh(\psi_1 + \psi_2)$, of $c_{12} = c_1c_2 + s_1s_2$ and $s_{12} = s_1c_2 + c_1s_2$ lead to

$$[\mathbf{L}(\psi_1)][\mathbf{L}(\psi_2)] = \begin{pmatrix} c_{12} & s_{12}\\ s_{12} & c_{12}\end{pmatrix} = [\mathbf{L}(\psi_1 + \psi_2)],$$

which is the result we set out to show.

(2.9) **Transformation multiplication leads to velocity addition rule** With the identification of (2.59) $\beta = -\tanh\psi$

$$\frac{u}{c} = \beta_1 = -\tanh\psi_1, \quad -\frac{v}{c} = \beta_2 = -\tanh\psi_2$$

and the transformation multiplication (2.79)

$$\frac{u'}{c} = \beta_{12} = -\tanh\psi_{12} = -\tanh(\psi_1 + \psi_2),$$

the velocity addition rule (2.24) follows from the trigonometric identity of

$$\tanh(\psi_1 \pm \psi_2) = \frac{\tanh\psi_1 \pm \tanh\psi_2}{1 \pm \tanh\psi_1\tanh\psi_2}.$$

(2.10) **Spacetime diagram depicting relativity of simultaneity**

(a) **The diagram for the observer at rest**: The two lightening bolts taking place simultaneously (say at $t = 0$), their worldpoints lie equidistant from the origin (the observer). The observer's worldline being the vertical time-axis, and the light coming from the lightnings trace out two 45° null worldlines, one with positive and other with negative slope; they meet the vertical worldline of the observer at one point.

(b) **The diagram for the moving observer**: The new space and time axes now "close in" as in Fig. 2.10. The worldline for the observer is now the tilted new time axis. The light worldlines are still the **same** two 45° lines as in (a). Consequently they will meet the observer's worldline (the new time axis) at two different points. They will no longer be perceived to be simultaneous.

(2.11) **Length contraction and light-pulse clock** In the rest frame of the clock, the total time $\triangle t'$ for a light pulse to go from one end to another, and back, is the sum $\triangle t' = \triangle t'_1 + \triangle t'_2$ where $\triangle t'_2$ is the time for the pulse to make the return trip. Clearly $\triangle t'_1 = \triangle t'_2 = L'/c$, where L' is the rest frame length of this clock. Now consider the clock in motion, moving (with velocity v) from left to right. The path the pulse must travel is lengthened when going from left to right, and shortened on the return trip, by the fact that the ends are moving to the right:

$$c\triangle t_1 = L + v\triangle t_1, \quad c\triangle t_2 = L - v\triangle t_2,$$

where L and $\triangle t$ are the length and time measured in the moving frame,

$$\triangle t = \triangle t_1 + \triangle t_2 = \frac{L}{c-v} + \frac{L}{c+v} = \gamma^2 \frac{2L}{c}. \tag{C.4}$$

Using the time dilation formula, we can compare

$$\triangle t = \gamma \triangle t' = \gamma \frac{2L'}{c}$$

to the result of (C.4) to obtain the Lorentz length contraction formula of $L = L'/\gamma$.

(2.12) **Pion decay-length in the laboratory** The naive calculation is incorrect because the half-life time $\tau_0 = 1.77 \times 10^{-8}$ s is the lifetime measured by a clock at rest with respect to the pion. The speed of $0.99c$ corresponds to $\gamma = 7.1$. (a) In the laboratory, the observer will see the pion decay time dilated to $\tau = \gamma\tau_0 = 7.1 \times 1.77 \times 10^{-8}$ s $= 1.26 \times 10^{-7}$ s, hence a decay length seven times longer, close to 38 m. (b) In the rest frame of the pion, this 38 m is viewed as having a contracted length of 5.3 m, which when divided by the particle speed of $0.99c$ yields its half-life time of $\tau_0 = 1.77 \times 10^{-8}$ s.

(2.14) Invariant spacetime interval and relativity of simultaneity

(a) Invariant spacetime interval:

$$-c^2\Delta t'^2 + \Delta x'^2 = \Delta x^2$$

or

$$c\Delta t' = \sqrt{\Delta x'^2 - \Delta x^2}.$$

(b) Lorentz transformation for the space coordinate $\Delta x' = \gamma \Delta x$ implies

$$\gamma = (1-\beta^2)^{-1/2} = \left(\frac{\Delta x'}{\Delta x}\right).$$

Hence

$$\gamma\beta = \sqrt{\gamma^2 - 1} \quad = \sqrt{\left(\frac{\Delta x'}{\Delta x}\right)^2 - 1}.$$

The Lorentz transformation for the time coordinates then leads to the same result as in (a),

$$c\Delta t' = \gamma\beta\Delta x = \sqrt{\Delta x'^2 - \Delta x^2}.$$

(2.15) More simultaneity calculations

(a) Given the Lorentz transformation

$$\Delta x' = \gamma(\Delta x - \beta c \Delta t), \tag{C.5}$$

$$c\Delta t' = \gamma(c\Delta t - \beta \Delta x), \tag{C.6}$$

and its inverse

$$\Delta x = \gamma(\Delta x' + \beta c \Delta t'), \tag{C.7}$$

$$c\Delta t = \gamma(c\Delta t' + \beta \Delta x'), \tag{C.8}$$

it is clear that $\Delta t' = 0$ implies, through (C.6),

$$\Delta t = \frac{\beta}{c}\Delta x,$$

through (C.8),

$$\Delta t = \frac{\beta}{c}\gamma\Delta x'. \tag{C.9}$$

These two equalities require the consistency condition:

$$\Delta x = \gamma\Delta x', \tag{C.10}$$

which is compatible with the Lorentz transformation (C.7) with $\Delta t' = 0$.

(b) Our derivation of length contraction in Section 2.4 would have us expecting this result of $\Delta x' = \gamma^{-1}\Delta x$ because the key input of the two ends of an object being measured at the same time in the "moving frame" is satisfied by our $\Delta t' = 0$ condition.

(c) In Section 2.2.2 we have shown that the time intervals, respectively from approaching bulb and receding bulb, for the light signals to reach the platform observer are

$$t_1 = \frac{L_p}{2c}\frac{1}{1+\beta}, \quad t_2 = \frac{L_p}{2c}\frac{1}{1-\beta},$$

where L_p is the length of the moving rail-car as seen by the platform observer. Because of length contraction, it should be $\gamma^{-1}\Delta x'$ as the rail car length seen by the O′ observer is $\Delta x'$. In this way we have the time difference

$$\Delta t = t_2 - t_1 = \frac{\beta}{c}\gamma^2 L_p = \frac{\beta}{c}\gamma\Delta x'$$

in agreement with (C.9). With respect to the O observer, the emission points are located at

$$x_1 = -ct_1 = -\frac{\Delta x'}{2\gamma}\frac{1}{1+\beta}, \quad x_2 = ct_2 = \frac{\Delta x'}{2\gamma}\frac{1}{1-\beta}.$$

Hence, according to the platform observer, the two emission events have a separation of

$$\Delta x = x_2 - x_1 = \frac{\Delta x'}{2\gamma}\left(\frac{1}{1-\beta} + \frac{1}{1+\beta}\right) = \gamma\Delta x',$$

which agrees with the result (C.10) obtained from Lorentz transformation.

(3.1) **Inclined plane, pendulum and EP**

(a) **Inclined plane:** The $F = ma$ equation along the inclined plane, is $m_I a = m_G g \sin\theta$, leading to a material-dependent acceleration:

$$a_A = g\sin\theta\left(\frac{m_G}{m_I}\right)_A.$$

(b) **Pendulum:** For the simple pendulum with a light string of length L, we have

$$m_I L\frac{d^2\theta}{dt^2} = -m_G g\sin\theta.$$

This has the form of a simple harmonic oscillator equation when approximated by $\sin\theta \approx \theta$, leading to a period of

$$T_A = \frac{2\pi}{\omega} = 2\pi\sqrt{\frac{L}{g}\left(\frac{m_I}{m_G}\right)_A}$$

for a blob made up of a particular material A.

(3.2) **Two EP brain-teasers**

(a) **Forward leaning balloon:** According to EP the effective gravity is the vector sum $\mathbf{g}_{\text{eff}} = \mathbf{g} + (-\mathbf{a})$, where $\mathbf{g}$ is the normal gravity (pointing vertically downward) while $\mathbf{a}$ is the acceleration of the vehicle. The buoyant force is always opposite to $\mathbf{g}_{\text{eff}}$.

(b) **A toy for Einstein:** What is normally difficult to do is to have a net force pulling the ball back into the bowl. The net force is the combination of gravity and spring restoring force. But the task can be made easy by dropping the whole contraption—because EP informs us that gravity would disappear in this freely falling system. Without the interference of gravity, the spring will pull back the ball each time without any difficulty.

(3.3) **Gravitational time dilation and the twin paradox** According to EP, the (de)acceleration needed to turn the spaceship around is equivalent to a gravitational field, which has an effect on the rate of time evolution. Now the clock on the turn-around rocketship is accelerating along with the spaceship, hence in "free fall" in this equivalent gravitational field (pointing toward the earth). According to the EP, this can be treated as an inertial observer. It can compare the outward-bound and inward-bound clocks in exactly the same way as the free-fall observer comparing the two clocks located at two different gravitational potential points as discussed in Section 3.3.1 in the paragraph with the sub-heading of "A more direct derivation." (Since this clock is just the onboard clock we have $\beta_1 = \beta_2 = 0$, $\tau_1 = t_1^{\mathrm{ff}}$, and $\tau_2 = t_2^{\mathrm{ff}}$ in (3.35).) The relative speed of the rocket ship before and after the turnaround has been computed in Section 3.4, (A.7), to be $\beta_{12} = \frac{40}{41}$, the SR time dilation formula, being applicable to this inertial frame, yields $t_1^{\mathrm{ff}} = t_2^{\mathrm{ff}}/\sqrt{1-\beta_{12}^2}$. Putting all these relations together we have the expected result of $\tau_1 = \frac{9}{41}\tau_2$.

(3.4) **The global position system**

(a) A satellite's centripetal acceleration is produced by earth's gravity:

$$\frac{v_s^2}{r_s} = G_{\mathrm{N}} \frac{M_\oplus}{r_s^2}.$$

The orbit period T_s is related to the radius and tangential velocity:

$$T_s = \frac{2\pi r_s}{v_s}.$$

Knowing that $T_s = 12$ h $= 4.32 \times 10^4$ s we can find r_s and v_s from these two equations:

$$r_s \simeq 2.7 \times 10^7 \text{ m} \simeq 4.2 R_\oplus, \quad v_s \simeq 3.9 \text{ km/s},$$

where $R_\oplus$ is earth's radius.

(b) The SR time dilation factor being $\gamma_s = (1-\beta_s^2)^{-1/2} = 1 + \beta_s^2/2 + \cdots$ the fractional change is then

$$\frac{1}{2}\beta_s^2 = \frac{1}{2}\left(\frac{v_s}{c}\right)^2 \simeq 0.85 \times 10^{-10}.$$

Here we have neglected the rotational speed of the clock on the ground—the corresponding β^2 value is a hundred times smaller even for the largest value on the equator.

(c) The gravitational time dilation effect is given by (3.31) with $\Phi = -G_{\rm N}M/r$:

$$\frac{\Phi_\oplus - \Phi_s}{c^2} = -G_{\rm N}\frac{M_\oplus}{c^2}\left(\frac{1}{R_\oplus} - \frac{1}{r_s}\right) \simeq -5.2 \times 10^{-10}.$$

Thus, the general relativity (GR) effect is about six times larger than the special relativity (SR) effect.

(d) In one minute duration

$$(\Delta t)_{\rm GR} \simeq -30 \text{ ns}, \quad (\Delta t)_{\rm SR} \simeq 5 \text{ ns}.$$

The gravitational effect makes the satellite clock go faster because it is at a higher gravitational potential. The SR dilation slows it down. The net effect is to make the clock in the satellite, when compared to the clock on the ground, run faster by about 25 ns for every passage of 1 min.

Here is an example of the practical application in our daily life of this "pure science" of general relativity.

(4.2) **Basis vectors on a spherical surface** The respective basis vectors are

$$\mathbf{e}_r = \hat{\mathbf{u}}_r, \quad \mathbf{e}_\phi = R\sin\frac{r}{R}\hat{\mathbf{u}}_\phi,$$

where $\hat{\mathbf{u}}_r$ is the unit vector in the radial direction, and $\hat{\mathbf{u}}_\phi$ is perpendicular to $\hat{\mathbf{u}}_r$ in the "tangential" direction. The resultant metric matrix, according to Eq. (4.7), is

$$g_{ab} = \begin{pmatrix} \mathbf{e}_r \cdot \mathbf{e}_r & \mathbf{e}_r \cdot \mathbf{e}_\phi \\ \mathbf{e}_\phi \cdot \mathbf{e}_r & \mathbf{e}_\phi \cdot \mathbf{e}_\phi \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & R^2\sin^2(r/R) \end{pmatrix}.$$

(4.3) **Coordinate transformation of the metric** Given the transformation (4.18), we have the inverse matrix

$$[\bar{\mathbf{R}}] = \begin{pmatrix} (\cos(r/R))^{-1} & 0 \\ 0 & 1 \end{pmatrix}.$$

Use Eq. (2.45) to act on the metric in the polar system to obtain that for the cylindrical system:

$$\begin{aligned}[\bar{\mathbf{R}}^{\sf T}][\mathbf{g}][\bar{\mathbf{R}}] &= [\bar{\mathbf{R}}^{\sf T}]\begin{pmatrix} 1 & 0 \\ 0 & R^2\sin^2(r/R) \end{pmatrix}[\bar{\mathbf{R}}] \\ &= \begin{pmatrix} (\cos^2(r/R))^{-1} & 0 \\ 0 & R^2\sin^2(r/R) \end{pmatrix} \\ &= \begin{pmatrix} [1-(\rho^2/R^2)]^{-1} & 0 \\ 0 & \rho^2 \end{pmatrix} = [\mathbf{g}'].\end{aligned}$$

(4.4) **Geodesics on simple surfaces**

(a) **Flat plane:** For this 2D space with Cartesian coordinates $(x^1, x^2) = (x, y)$, the metric is $g_{ab} = \delta_{ab}$. The second term in the geodesic Eq. (4.30) vanishes, as well as the two components of the equation $d\dot{x}^\nu/d\sigma$

$$\ddot{x} = 0 \quad \text{and} \quad \ddot{y} = 0,$$

which have respective solutions of

$$x = \mathrm{A} + \mathrm{B}\sigma \quad \text{and} \quad y = \mathrm{C} + \mathrm{D}\sigma.$$

They can be combined as

$$y = \alpha + \beta x$$

with (A, B, C, D) and (α, β) being constants. We recognize this as the equation for a straight line.

(b) **Spherical surface:** For a 2-sphere, we choose the coordinates $(x^1, x^2) = (\theta, \phi)$ with a diagonal metric similar to (4.13) with elements $g_{\theta\theta} = R^2$ and $g_{\phi\phi} = R^2 \sin^2\theta$. For the θ component of the geodesic Eq. (4.30) is

$$\ddot{\theta} = \sin\theta\cos\theta\dot{\theta}\dot{\phi}^2, \tag{C.11}$$

the ϕ component equation,

$$2\sin\theta\cos\theta\dot{\theta}\dot{\phi} + \sin^2\theta\ddot{\phi} = 0. \tag{C.12}$$

Instead of working out the full parametric representation, we will just check that $\phi =$ constant and $\theta = \alpha + \beta\sigma$ solve these two equations. Clearly these solutions describe longitudinal great circles on the sphere.

(4.5) **Locally flat metric** The distance between two neighboring points can be rearranged by $\pm(g_{12}dx^2)^2/g_{11}$:

$$\begin{aligned} ds^2 &= g_{11}(dx^1)^2 + 2g_{12}dx^1dx^2 + g_{22}(dx^2)^2 \\ &= \left(\sqrt{g_{11}}dx^1 + \frac{g_{12}dx^2}{\sqrt{g_{11}}}\right)^2 + \left(g_{22} - \frac{g_{12}^2}{g_{11}}\right)(dx^2)^2. \end{aligned}$$

The new coordinate system $(\bar{x}^1, \bar{x}^2)$ has the metric $\bar{g}_{ab} = \delta_{ab}$ because $ds^2 = (d\bar{x}^1)^2 + (d\bar{x}^2)^2$ where

$$d\bar{x}^1 = \sqrt{g_{11}}dx^1 + \frac{g_{12}dx^2}{\sqrt{g_{11}}}, \quad d\bar{x}^2 = \sqrt{g_{22} - \frac{g_{12}^2}{g_{11}}}dx^2.$$

On the other hand, if the original metric determinant is negative, $g_{11}g_{22} - g_{12}^2 < 0$, then $ds^2 = (d\bar{x}^1)^2 - (d\bar{x}^2)^2$ with

$$d\bar{x}^2 = \sqrt{\frac{g_{12}^2}{g_{11}} - g_{22}dx^2}.$$

(4.7) **3-sphere and 3-pseudosphere**

(a) **3D flat space**

$$x = r\sin\theta\cos\phi, \quad y = r\sin\theta\sin\phi, \quad z = r\cos\theta.$$

The relation for the solid angle factor follows simply from the two expressions for the invariant separations in two coordinate systems:

$$ds^2 = dx^2 + dy^2 + dz^2 = dr^2 + r^2 d\Omega^2. \qquad \text{(C.13)}$$

(b) **3-sphere** Given the metric for 3-sphere being

$$ds^2 = dr^2 + \left(R \sin \frac{r}{R}\right)^2 d\Omega^2. \qquad \text{(C.14)}$$

The relation from part (a) $r^2 d\Omega^2 = dx^2 + dy^2 + dz^2 - dr^2$ suggests

$$\left(R \sin \frac{r}{R}\right)^2 d\Omega^2 = dX^2 + dY^2 + dZ^2 - d\left(R \sin \frac{r}{R}\right)^2.$$

Substituting this into (C.14), we have

$$ds^2 = dW^2 + dX^2 + dY^2 + dZ^2$$

if we identify $dW = \sin(r/R)dr$. This invariant interval implies a Euclidian metric $g_{\mu\nu} = \mathrm{diag}(1, 1, 1, 1)$. Also, we have

$$W = R\cos\frac{r}{R}, \quad X = \left(R\sin\frac{r}{R}\right)\sin\theta\cos\phi,$$
$$Y = \left(R\sin\frac{r}{R}\right)\sin\theta\sin\phi, \quad Z = \left(R\sin\frac{r}{R}\right)\cos\theta.$$

This set of relations lead immediately to the constraint $W^2 + X^2 + Y^2 + Z^2 = R^2$.

(c) **3-pseudosphere** With $W = R\cosh(r/R)$, the relations

$$W = R\cosh\frac{r}{R}, \quad X = \left(R\sinh\frac{r}{R}\right)\sin\theta\cos\phi,$$
$$Y = \left(R\sinh\frac{r}{R}\right)\sin\theta\sin\phi, \quad Z = \left(R\sinh\frac{r}{R}\right)\cos\theta,$$

lead, through the trigonometric relation $\cosh^2\chi - \sinh^2\chi = 1$ to

$$ds^2 = -dW^2 + dX^2 + dY^2 + dZ^2$$

thus a Minkowski metric $\eta_{\mu\nu} = \mathrm{diag}(-1, 1, 1, 1)$ and the condition

$$-W^2 + X^2 + Y^2 + Z^2 = -R^2.$$

(4.8) **Volume of higher dimensional space**

$$dV = \sqrt{\det g}\prod_i dx^i. \qquad \text{(C.15)}$$

(a) **3D flat space** For Cartesian coordinates $\sqrt{\det g} = 1$, (C.15) reduces to $dV = dx\,dy\,dz$, and for spherical coordinates $\sqrt{\det g} = r^2\sin\theta$ and $dV = r^2\sin\theta dr\,d\theta\,d\phi$.

(b) **3-sphere** From (C.14) we have $\sqrt{\det g} = R^2\sin^2(r/R)\sin\theta$, thus the volume of a 3-sphere with radius R can be calculated:

$$R^2\int_0^{\pi R}\sin^2\frac{r}{R}dr\int_0^{\pi}\sin\theta d\theta\int_0^{2\pi}d\phi$$

to be $2\pi^2 R^3$.

(4.9) **Non-Euclidean relation between radius and circumference of a circle**

(a) **The case of a sphere**: The radius of a circle being the displacement ds along a constant radial coordinate ($dr = 0$), we have from either (4.40) or (4.42), $ds = R\sin(r/R)d\phi$. Thus, making a Taylor series expansion of the circumference $S = \int ds$, we have:

$$S = 2\pi R \sin\frac{r}{R} = 2\pi R\left(\frac{r}{R} - \frac{1}{3!}\frac{r^3}{R^3} + \cdots\right)$$
$$= 2\pi r - \frac{1}{R^2}\frac{\pi r^3}{3} + \cdots$$

which is just the claimed result in (4.49) with $K = 1/R^2$.

(b) **The case of a pseudosphere**: For $k = -1$ surface, the displacement, according to either (4.40) or (4.42), is given by $ds = R\sinh(r/R)d\phi$, giving a circumference of $S = 2\pi R\sinh(r/R)$. Since the Taylor expansion of the hyperbolic sine differs from that for the sine function in the sign of the cubic term, again we obtain the result in agreement with (4.49) with $K = -1/R^2$. Thus, on a pseudospherical surface, the circumference of a circle with radius r is $S > 2\pi r$.

(4.10) **Angular excess and polygon area**

Any polygon is made up of triangles.

(5.2) **Spatial distance and spacetime metric**

The spacetime separation vanishes ($ds^2 = 0$) for a light pulse:

$$g_{00}(dx^0)^2 + 2g_{0i}\,dx^i dx^0 + g_{ij}\,dx^i dx^j = 0.$$

Solving this quadratic equation for the coordinate time interval that takes the pulse going from A to B

$$dx^0_{\rm AB} = -\frac{g_{0i}dx^i}{g_{00}} - \frac{\sqrt{(g_{0i}g_{0j} - g_{00}g_{ij})dx^i dx^j}}{g_{00}}$$

and the time for it to go from B to A (involving the change of $dx^i \to -dx^i$)

$$dx^0_{\rm BA} = +\frac{g_{0i}dx^i}{g_{00}} - \frac{\sqrt{(g_{0i}g_{0j} - g_{00}g_{ij})dx^i dx^j}}{g_{00}}.$$

Therefore the total coordinate time

$$dx^0 = dx^0_{\rm AB} + dx^0_{\rm BA},$$

which is related to the proper time interval $d\tau_{\rm A}$ (cf. Problem 5.1), hence the spatial distance dl,

$$dl \equiv \frac{cd\tau_{\rm A}}{2} = \frac{\sqrt{-g_{00}}dx^0}{2} = \sqrt{\left(g_{ij} - \frac{g_{0i}g_{0j}}{g_{00}}\right)dx^i dx^j}.$$

Namely

$$\gamma_{ij} = g_{ij} - \frac{g_{0i}g_{0j}}{g_{00}}.$$

Thus $\gamma_{ij} \neq g_{ij}$ when $g_{0i} \neq 0$.

(5.3) **Non-Euclidean geometry of a rotating cylinder** Let us denote the spatial coordinates as follows:

$$(ct, r, \phi, z) \quad \text{lab observer,}$$
$$(ct, r_0, \phi_0, z) \quad \text{observer on the rotating disc.}$$

They are related by (cf. Fig. 5.1)

$$r = r_0, \quad \phi = \phi_0 + \omega t.$$

We shall ignore the vertical coordinate z below.

The invariant separation written in terms coordinates at rest with respect to the observer on the rotating disc is (cf. (4.33))

$$ds^2 = -c^2dt^2 + dr_0^2 + r_0^2 d\phi_0^2,$$

which can be written in terms of the lab coordinate (cf. (Cook, 2004)) by substituting in $d\phi_0 = d\phi - \omega dt$:

$$ds^2 = -\left[1 - \left(\frac{\omega r}{c}\right)^2\right] c^2dt^2 + dr^2 + r^2 d\phi^2 - 2\omega r^2 dt\, d\phi.$$

Namely the metric with respect to the (ct, r, ϕ) coordinate has elements

$$g_{00} = -\left[1 - \left(\frac{\omega r}{c}\right)^2\right], \quad g_{rr} = 1, \quad g_{\phi\phi} = r^2, \quad g_{0\phi} = -\frac{\omega r^2}{c}.$$

From Problem 5.2, we have the spatial distance

$$dl^2 = \left(g_{ij} - \frac{g_{0i}g_{0j}}{g_{00}}\right) dx^i dx^j = dr^2 + \frac{r^2 d\phi^2}{1 - (\omega r/c)^2}$$

showing clearly length contraction of the circumference, but not the radius.

(5.5) **The geodesic equation and light deflection** The curve parameter σ can be taken to be the proper time τ. The geodesic Eq. (5.9)

$$\frac{d}{d\tau}\frac{dx^\mu}{d\tau} + \Gamma^\mu_{\nu\lambda}\frac{dx^\nu}{d\tau}\frac{dx^\lambda}{d\tau} = 0,$$

which, after using $p^\mu \propto dx^\mu/d\tau$, can be written as

$$\frac{d}{d\tau}p^\mu + \Gamma^\mu_{\nu\lambda}p^\nu\frac{dx^\lambda}{d\tau} = 0$$

or equivalently

$$dp^\mu = -\Gamma^\mu_{\nu\lambda}p^\nu dx^\lambda.$$

We are interested in the $\mu = 2$ component

$$\begin{aligned} dp_y &= -\Gamma^2_{00}p^0dx^0 - \Gamma^2_{11}p^1dx^1 - \Gamma^2_{10}p^1dx^0 - \Gamma^2_{01}p^0dx^1 \\ &= -(\Gamma^2_{00} + \Gamma^2_{11} + 2\Gamma^2_{10})pdx, \end{aligned} \tag{C.16}$$

where we have used $dx^\mu = (dx, dx, 0, 0)$ and $p^\mu = (p, p, 0, 0)$. Christoffel symbols can be calculated by (5.10). Since we are working in the weak-field approximation, that is, the metric is very close to being the flat space Minkowski metric $\eta_{\mu\nu} = \text{diag}\,(-1, 1, 1, 1)$, and the Christoffel symbols (being the derivatives of the metric) must also be small. Thus the metric on the left-hand side (LHS) of (5.10) can be taken to be $\eta_{\mu\nu}$ and can be "moved" to the right-hand side (RHS) by contracting both sides by $\eta_{\lambda\rho}$ and using $\eta_{\lambda\rho}\eta_{\rho\sigma} = \delta_{\lambda\sigma}$,

$$\Gamma^2_{\mu\nu} = \frac{1}{2}\eta_{2\rho}\left[\frac{\partial g_{\mu\rho}}{\partial x^\nu} + \frac{\partial g_{\nu\rho}}{\partial x^\mu} - \frac{\partial g_{\mu\nu}}{\partial x^\rho}\right]$$
$$= \frac{1}{2}\left[\frac{\partial g_{\mu 2}}{\partial x^\nu} + \frac{\partial g_{\nu 2}}{\partial x^\mu} - \frac{\partial g_{\mu\nu}}{\partial x^2}\right].$$

If the only position-dependent metric element is $g_{00} = -1 - \Phi/c^2$ (as suggested by EP physics) and thus the only nonzero term on the RHS of (C.16) is

$$\Gamma^2_{00} = \frac{-1}{c^2}\frac{\partial\Phi}{\partial y}.$$

This way we get

$$\delta_{\mathrm{EP}} = \int \frac{dp_y}{p} = \frac{1}{c^2}\int \frac{\partial\Phi}{\partial y}dx \tag{C.17}$$

which is the result obtained by Huygens' principle in Eq. (3.44). For the argument that the GR value is twice that of the EP value, see Section 6.2.1.

(5.7) **The matrix for tidal forces is traceless** We can take the trace of the tidal force matrix as

$$\delta_{ij}\frac{\partial^2\Phi}{\partial x^i \partial x^j} = \frac{\partial}{\partial x^i}\frac{\partial}{\partial x^i}\Phi = \nabla^2\Phi.$$

Since the mass density vanishes ($\rho = 0$) at any field point away from the source, the Newtonian field Eq. (5.5) informs us that the gravitational potential satisfies the Laplace equation $\nabla^2\Phi = 0$.

(5.8) **G_{N} as a conversion factor** One easily finds that this yields the dimension relation (curvature) = (length)$^{-2}$. This is consistent with the fact that curvature is the second derivative of the metric, which is dimensionless.

(6.1) **Energy relation for a particle moving in the Schwarzschild spacetime**. The $r^* = 0$ limit (6.40) is

$$-c^2\left(\frac{dt}{d\tau}\right)^2 + \left(\frac{dr}{d\tau}\right)^2 + r^2\left(\frac{d\phi}{d\tau}\right)^2 = -c^2,$$

where τ is the proper time $d/d\tau = \gamma d/dt$ with $\gamma = (1-v^2/c^2)^{-1/2}$. Multiplying by a factor of m^2c^2 on both sides, we obtain

$$\gamma^2 m^2[c^4 - c^2v^2] = m^2c^4,$$

where $v^2 = (dr/dt)^2 + r^2(d\phi/dt)^2$ is the velocity (squared) in the spherical coordinate system (r, θ, ϕ) when the polar angle θ is fixed. We recognize this is the energy–momentum relation $E^2 = p^2c^2 + m^2c^4$ after identifying the relativistic expression for energy $E = \gamma mc^2$ and momentum $p = \gamma mv$.

(6.2) **Equation for a light trajectory** For a lightlike worldline $ds^2 = 0$ the Lagrangian $L = ds^2/d\sigma^2$ must also vanish. Following the same steps as (6.44) to (6.50), we have

$$\left(\frac{dr}{d\sigma}\right)^2+\left(1-\frac{r^*}{r}\right)\frac{\lambda^2}{4r^2}=c^2\eta^2.$$

and, after the usual change of variables,

$$u''+u-\epsilon u^2=0.$$

For a perturbative solution of $u=u_0+\epsilon u_1$,

$$(u_0''+u_0)+\epsilon(u_1''+u_1-u_0^2)+\cdots=0.$$

The zeroth order, being a "simple harmonic oscillator" equation, has the solution $u_0=r_{\min}^{-1}\sin\phi$. To solve the first-order equation

$$\frac{d^2u_1}{d\phi^2}+u_1=\frac{1-\cos 2\phi}{2r_{\min}^2}$$

one tries $u_1=\alpha+\beta\cos 2\phi$, and finds $\alpha=(2r_{\min}^2)^{-1}$ and $\beta=(6r_{\min}^2)^{-1}$. Putting the zeroth and first order terms together we get

$$\frac{1}{r}=\frac{\sin\phi}{r_{\min}}+\frac{3+\cos 2\phi}{4}\frac{r^*}{r_{\min}^2}.$$

In the absence of gravity ($r^*=0$), the asymptotes ($r=\mp\infty$) correspond to $\phi_{-\infty}=\pi$ and $\phi_{+\infty}=0$, and the trajectory is straight line (no deflection). When gravity is turned on, there is an angular deflection $\delta=(\phi_{-\infty}-\phi_{+\infty}-\pi)$. Picking our coordinates so that $\phi_{-\infty}=\pi+\delta/2$ and $\phi_{+\infty}=-\delta/2$ and the trajectory equations yields

$$0=-\sin\frac{\delta}{2}+\frac{3+\cos\delta}{4}\frac{r^*}{r_{\min}}.$$

For small deflection angle δ,

$$0=-\frac{\delta}{2}+\frac{r^*}{r_{\min}}.$$

Thus the result $\delta_{\mathrm{GR}}=2r^*/r_{\min}$.

(6.5) **Circular orbits** For a circular orbit, the radial distance and the orbital angular momentum must satisfy a definite relation so that the effective potential is minimized (at this radial distance):

$$\frac{\partial\Phi_{\mathrm{eff}}}{\partial r}=0,$$

which turns out to be

$$l^2=G_{\mathrm{N}}Mm^2r\left(1-\frac{3}{2}\frac{r^*}{r}\right)^{-1}. \qquad \text{(C.18)}$$

Furthermore, the total energy must be just equal to the potential energy:

$$\mathcal{K}=m\Phi_{\mathrm{eff}}$$

or, using the suggested form for $\mathcal{K}$ and Φ_{eff}, it may be written as

$$\frac{\eta^2-1}{2}=\frac{1}{2}\left[\left(1-\frac{r^*}{r}\right)\left(1+\frac{l^2}{m^2r^2c^2}\right)-1\right].$$

After plugging the result in (C.18), one finds

$$\eta^2 = \left(1 - \frac{r^*}{r}\right)^2 \left(1 - \frac{3}{2}\frac{r^*}{r}\right)^{-1}. \tag{C.19}$$

(7.2) **Luminosity distance to the nearest star** The observed flux being $f = L/4\pi d^2$, we have

$$d_* = \left(\frac{f_\odot}{f_*}\right)^{1/2} \times \text{AU} = 3 \times 10^5 \text{AU} = 1.5 \text{ pc}.$$

(7.3) **Gravitational frequency shift contribution to the Hubble redshift** Given that the gravitational redshift is given by (3.26), the redshift the photon suffers to overcome the gravitational pull of a galaxy, which has a mass $M_\text{G} = O(10^{11} M_\odot)$ with a linear dimension $R_\text{G} = O(10^{12} R_\odot)$, can be estimated to be

$$z_\text{G} = \frac{M_\text{G}}{M_\odot}\frac{R_\odot}{R_\text{G}} z_\odot = O(10^{-7}),$$

where we have approximated the galaxy as a spherical system and used solar redshift $z_\odot = O(10^{-6})$. Thus, the shift due to gravity is quite negligible.

(7.4) **Energy content due to star light** Let us denote the average stellar luminosity by L_* and star number density by n. Their product is then the luminosity density as given by (7.21),

$$nL_* = 0.2 \times 10^9 \frac{L_\odot}{(\text{Mpc})^3} = 2.6 \times 10^{-33} \text{ W m}^{-3},$$

which is the energy emitted per unit volume per unit time. Stars have been assumed to be emitting light at this luminosity during the entire $t_0 \simeq t_\text{H} \simeq 13.5$ Gyr $= 4.3 \times 10^{17}$ s, leading to an energy density contribution at present of

$$\rho_* c^2 = nL_* t_\text{H} \simeq 10^{-15} \text{ J m}^{-3}$$

or, using (7.19), a density ratio

$$\Omega_* = \frac{\rho_*}{\rho_\text{c}} \simeq 10^{-6}.$$

(7.5) **Night sky as bright as day** Flux being the watts per unit area, the total flux due to all the starlights is, according to (7.2),

$$f_* = nL_* cH_0^{-1} \simeq \left(0.2 \times 10^9 \frac{L_\odot}{\text{Mpc}^3}\right)(cH_0^{-1})$$
$$= 0.8 \times 10^{12} \frac{L_\odot}{\text{Mpc}^2} = 2.5 \times 10^{-10} \frac{L_\odot}{4\pi(\text{AU})^2}.$$

Thus, we need to lengthen the age by a factor of 4 billion before we can get a night sky as bright as day!

(7.6) **The Virial theorem** Differentiating $G \equiv \sum_n \mathbf{p}_n \cdot \mathbf{r}_n$ with respect to time

$$\frac{dG}{dt} = \sum\nolimits_n \mathbf{F}_n \cdot \mathbf{r}_n + \sum\nolimits_n \mathbf{p}_n \cdot \dot{\mathbf{r}}_n$$

and taking the time average on both sides, we have the LHS

$$\left\langle \frac{dG}{dt} \right\rangle = \lim_{T\to\infty} \frac{1}{T} \int_0^T \frac{dG}{dt} dt = \lim_{T\to\infty} \frac{G(T) - G(0)}{T},$$

which vanishes because this is a bound system. We then have, with $\sum_n \mathbf{p}_n \cdot \dot{\mathbf{r}}_n = \sum mv^2 = 2T$,

$$2\langle T \rangle = -\left\langle \sum_n \mathbf{F}_n \cdot \mathbf{r}_n \right\rangle = \left\langle \sum_n \nabla V_n \cdot \mathbf{r}_n \right\rangle = \left\langle \sum_n \frac{\partial V_n}{\partial r_{ni}} r_{ni} \right\rangle$$

for inverse square force $V_n \propto r_n^{-1}$, thus

$$\frac{\partial V_n}{\partial r_{ni}} r_{ni} = -V_n$$

and finally the virial theorem result of $2\langle T \rangle = -\langle V \rangle$.

(7.8) **Wavelength in an expanding universe** A radial light signal follows the null worldline in the RW geometry and its proper distance is given by (7.49). Consider two successive wavecrests with wavelength λ; the second one is emitted (and observed) later by a time interval $\delta t = \lambda/c$. Both wavecrests travel the same distance $d_{\mathrm{p}}(\xi, t_0)$:

$$d_{\mathrm{p}} = \int_{t_{\mathrm{em}}}^{t_0} \frac{cdt}{a(t)} = \int_{t_{\mathrm{em}}+\lambda_{\mathrm{em}}/c}^{t_0+\lambda_0/c} \frac{cdt}{a(t)}.$$

After cancelling out the common interval of $(t_{\mathrm{em}} + \lambda_{\mathrm{em}}/c, t_0)$ from both sides of the integral equality, we have

$$\int_{t_{\mathrm{em}}}^{t_{\mathrm{em}}+\lambda_{\mathrm{em}}/c} \frac{cdt}{a(t)} = \int_{t_0}^{t_0+\lambda_0/c} \frac{cdt}{a(t)}.$$

Since the scale factor would not have changed much during the small time interval between these two crests

$$\frac{1}{a(t_{\mathrm{em}})} \int_{t_{\mathrm{em}}}^{t_{\mathrm{em}}+\lambda_{\mathrm{em}}/c} dt = \frac{1}{a(t_0)} \int_{t_0}^{t_0+\lambda_0/c} dt$$

which immediately leads to the expected result of

$$\frac{\lambda_0}{\lambda_{\mathrm{em}}} = \frac{a(t_0)}{a(t_{\mathrm{em}})}.$$

(7.9) **The steady-state universe**

(a) “Perfect CP” means that the universe is not only homogeneous in space but also in time.

(b) From (7.48) we have

$$\frac{da}{dt} = H_0 a$$

which has the solution $a(t) = \exp[H_0(t - t_0)]$. Thus $\dot{a} = H_0 a$ and $\ddot{a} = H_0^2 a$ so that $q_0 = -1$.

(c) According to (4.38), the curvature for the 3D space in the Steady-State Universe (SSU) is $K = kR^{-2}(t)$. Since the scale factor does depend on t, an unchanging K can come about only for the curvature signature $k = 0$. Namely, an SSU requires a 3D space with a flat geometry.

(d) The rate of mass creation per unit volume for a constant density

$$\frac{\dot{M}}{V} = \rho_{\mathrm{M}} \frac{\dot{V}}{V} = 3H_0\rho_{\mathrm{M}} \simeq 0.7 \times 10^{-24} \text{ g/year/km}^3.$$

Given that $m_{\mathrm{p}} = 1.7 \times 10^{-24}$ g, this means the creation of one hydrogen atom, in a cubic kilometer volume, every 2–3 years.

(7.10) The deceleration parameter and Taylor expansion of the scale factor

$$\begin{aligned} a(t) &\simeq a(t_0) + (t - t_0)\dot{a}(t_0) + \frac{1}{2}(t - t_0)^2\ddot{a}(t_0) \\ &= 1 + (t - t_0)H_0 - \frac{1}{2}(t - t_0)^2 q_0 H_0^2 \end{aligned} \tag{C.20}$$

and

$$\frac{1}{a(t)} \simeq 1 - (t - t_0)H_0 + (t - t_0)^2\left(1 + \frac{q_0}{2}\right)H_0^2. \tag{C.21}$$

(7.11) z^2 correction to the Hubble relation

(a) From (7.49)

$$d_{\mathrm{p}}(t_0) = a(t_0)\int_{t_{\mathrm{em}}}^{t_0} \frac{c\,dt}{a(t)}$$

and the first two terms of the Taylor series (C.21) we have

$$d_{\mathrm{p}}(t_0) = c(t_0 - t_{\mathrm{em}}) + \frac{c}{2}H_0(t_0 - t_{\mathrm{em}})^2. \tag{C.22}$$

The first term on the RHS is just the distance traversed by a light signal in a static environment; the second term represents the correction due to the expansion of the universe.

(b) $(t_0 - t_{\mathrm{em}})$ can be related to the redshift z through (7.54) and (C.21).

$$z = -1 + \frac{1}{a(t_{\mathrm{em}})} = (t_0 - t_{\mathrm{em}})H_0\left[1 + (t_0 - t_{\mathrm{em}})H_0\left(1 + \frac{q_0}{2}\right)\right]. \tag{C.23}$$

(c) Equation (C.23) can be inverted to yield

$$t_0 - t_{\rm em} \simeq \frac{z}{H_0}\left[1 - (t_0 - t_{\rm em})H_0\left(1 + \frac{q_0}{2}\right)\right]$$
$$\simeq \frac{z}{H_0}\left[1 - z\left(1 + \frac{q_0}{2}\right)\right]. \qquad \text{(C.24)}$$

Plug this expression for the look-back time into (C.22), we have

$$D_{\rm p}(t_0) \simeq \left[\frac{cz}{H_0} - \frac{cz^2}{H_0}\left(1 + \frac{q_0}{2}\right)\right] + \frac{cz^2}{2H_0}$$
$$= \frac{cz}{H_0}\left(1 - \frac{1+q_0}{2}z\right).$$

(8.2) **Newtonian interpretation of second Friedmann equation** For the pressureless matter used for our Newtonian system, cf. Fig. 8.1, the gravitational attraction by the whole sphere being $-G_{\rm N}M/r^2 = \ddot{r}$, or

$$-\frac{4\pi G_{\rm N} a\rho}{3} = \ddot{a}$$

which is just Eq. (8.2) without the pressure term.

(8.4) **The empty universe** The nontrivial solution to

$$\dot{a}^2 = -\frac{kc^2}{R_0^2}$$

is a negatively curved open universe $k = -1$ and, with $t_0 = c^{-1}R_0$,

$$a = \frac{t}{t_0}$$

which is just the straight-line $a(t)$ in Fig. 8.2. From (7.49) we can obtain the proper distance in terms of z.

$$d_{\rm p}(t_0) = \int_{t_{\rm em}}^{t_0} \frac{cdt}{a(t)} = ct_0 \int_{t_{\rm em}}^{t_0} t^{-1}dt = ct_0 \ln\left(\frac{t_0}{t_{\rm em}}\right) = ct_0 \ln(1+z),$$

where we have used $t_0/t_{\rm em} = (a(t_{\rm em}))^{-1} = (1+z)$. It is clear that for small redshift this equation reduces to the Hubble relation (7.5) with $H_0 = t_0^{-1}$. Namely, in an empty universe the age is given by the Hubble time $t_0 = t_{\rm H}$, and the "radius" by the Hubble length $R_0 = l_{\rm H} = ct_{\rm H}$.

(8.5) **Hubble plot in matter-dominated flat universe** Since distance modulus is a simple logarithmic expression (7.66) of luminosity distance $d_{\rm L}$, which in turn is related to our proper distance $d_{\rm p}$ by $d_{\rm L} = (1+z)d_{\rm p}$ of (7.61), all we need is to calculate the proper distance according to (7.49). This integration can be performed for this matter dominated flat universe, which has $a(t) = (t/t_0)^{2/3}$ as given in (8.30).

$$d_\mathrm{p}(t_0) = \int_{t_\mathrm{em}}^{t_0} \frac{cdt}{a(t)} = ct_0^{2/3}\int_{t_\mathrm{em}}^{t_0} t^{-2/3}dt = 3ct_0\left[1-\left(\frac{t_\mathrm{em}}{t_0}\right)^{1/3}\right]$$
$$= 3ct_0\left(1-[a(t_\mathrm{em})]^{1/2}\right) = \frac{2c}{H_0}(1-[1+z]^{-1/2}),$$

where we have used $a(t_\mathrm{em}) = (t_\mathrm{em}/t_0)^{2/3}$ one more time as well as the basic redshift relation of (7.54) and the age of a flat MDU $t_0 = \frac{2}{3}H_0^{-1}$ of (8.30). This way one finds the distance modulus to be

$$m - M = 5\log_{10}\frac{2cH_0^{-1}(1+z-[1+z]^{1/2})}{10\ \mathrm{pc}}.$$

(8.7) **Time and redshift of a light emitter** Plug (8.27), $a(t) = (t/t_0)^x$ into Eq. (7.49)

$$d_\mathrm{p}(t_0) = \int_{t_\mathrm{em}}^{t_0}\frac{cdt'}{a(t')} = \frac{ct_0}{1-x}\left[1-\left(\frac{t_\mathrm{em}}{t_0}\right)^{1-x}\right]. \qquad \text{(C.25)}$$

We can also express the time of light emission from a receding galaxy with redshift z. Through (7.54) and (8.27) we obtain

$$1+z = \frac{a(t_0)}{a(t_\mathrm{em})} = \left(\frac{t_0}{t_\mathrm{em}}\right)^x$$

or

$$t_\mathrm{em} = \frac{t_0}{(1+z)^{1/x}}. \qquad \text{(C.26)}$$

Plugging into (C.25), we have

$$d_\mathrm{p}(t_0) = \frac{ct_0}{1-x}\left[1-\frac{1}{(1+z)^{(1-x)/x}}\right]. \qquad \text{(C.27)}$$

We note in particular for a matter-dominated flat universe $x = \frac{2}{3}$ we have

$$d_\mathrm{p}(t_0) = 3ct_0\left[1-\frac{1}{(1+z)^{1/2}}\right], \qquad \text{(C.28)}$$

which agrees with the result obtained in Problem 8.5 and for a radiation-dominated flat universe $x = \frac{1}{2}$ we have

$$d_\mathrm{p}(t_0) = 2ct_0\left[1-\frac{1}{1+z}\right]. \qquad \text{(C.29)}$$

NB: These simple relations between redshift and time hold only for a universe with a single-component on energy content; moreover, it does not apply to the situation when the equation-of-state parameter is negative ($w = -1$), even though the energy content is a single-component case.

(8.8) **Scaling behavior of number density and Hubble's constant**

(a) For material particles the number density scales as the inverse volume factor.

$$\frac{n(t)}{n_0} = [a(t)]^{-3}.$$

The basic relation (7.54) between scale factor and redshift leads to

$$\frac{n(t)}{n_0} = (1+z)^3.$$

This scaling property also holds for radiation because $n \sim T^3 \sim a^{-3}$ as given in (8.35) and (8.40).

(b) We can obtain the scaling behavior of the Hubble parameter from Friedmann equation (8.1) for a flat universe:

$$\frac{\dot{a}^2}{a^2} = \frac{8\pi G_N \rho}{3}. \tag{C.30}$$

which can be written as

$$\frac{H^2}{H_0^2} = \frac{\rho}{\rho_{c,0}}. \tag{C.31}$$

For an epoch when the density is dominated by radiation $\rho \simeq \rho_R = \rho_{R,0} a^{-4}$, the above expression for H becomes

$$\frac{H^2}{H_0^2} = \Omega_{R,0}(1+z)^4. \tag{C.32}$$

In the entirely same way we can show that the Hubble parameter in a matter dominated epoch obeys

$$\frac{H^2}{H_0^2} = \Omega_{M,0}(1+z)^3. \tag{C.33}$$

(8.9) **Radiation and matter equality time** Since the universe from t_{RM} to t_γ is matter-dominated, we have from (8.30)

$$\frac{a(t_{RM})}{a(t_\gamma)} = \left(\frac{t_{RM}}{t_\gamma}\right)^{2/3}$$

or

$$t_{\mathrm{RM}} = \left[\frac{a(t_{\mathrm{RM}})}{a(t_\gamma)}\right]^{3/2} t_\gamma = \left[\frac{1+z_{\mathrm{RM}}}{1+z_\gamma}\right]^{-3/2} t_\gamma$$

$$\simeq \frac{t_\gamma}{10^{3/2}} \simeq 10,000 \text{ year.}$$

(8.10) **Density and deceleration parameter** Use the definition of w in (8.4), the second Friedmann Eq. (8.2) becomes

$$\frac{\ddot{a}(t)}{a(t)} = -\frac{4\pi G_{\mathrm{N}}}{3}\sum_i \rho_i(1+3w_i).$$

In terms of the deceleration parameter (7.67)

$$q_0 \equiv \frac{-\ddot{a}(t_0)}{a(t_0)H_0^2}$$

and the critical density (8.6)

$$\rho_{\mathrm{c},0} = \frac{3H_0^2}{8\pi G_{\mathrm{N}}}$$

the second derivative equation leads to the claimed result

$$q_0 = \frac{1}{2}\sum_i \Omega_{i,0}(1+3w_i) = \Omega_{\mathrm{R},0} + \frac{1}{2}\Omega_{\mathrm{M},0} + \cdots$$

(8.13) **Cosmological limit of neutrino mass** Even if we assume that all the nonbaryonic dark matter is made of three species (flavors) of neutrinos

$$\rho_{\mathrm{exotic}} = \sum_{i=1}^{3}\rho(\nu_i) = 3n_\nu\bar{m},$$

where n_ν is the neutrino number density and $\bar{m}$ is the average neutrino mass. From the neutrino and photon temperature of (8.75) and density being the cubic power of temperature (8.35),

$$n_\nu = \left(\frac{T_\nu}{T_\gamma}\right)^3 n_\gamma \simeq (1.7)^{-3} \times 400 \simeq 150 \text{ cm}^{-3}.$$

The energy density ratio becomes

$$\Omega_{\mathrm{exotic}} = \frac{3n_\nu\bar{m}c^2}{\rho_{\mathrm{c}}c^2} \simeq 0.26$$

Using the critical energy density value of (7.19), we have

$$\bar{m}c^2 \simeq \frac{0.26 \times 5,500}{3 \times 150} \simeq 3 \text{ eV}.$$

(8.14) **Temperature dipole anisotropy as Doppler effect** Recall that temperature scales as a^{-1}, that is, as inverse wavelength, or as frequency:

$$\frac{\delta T}{T} = \frac{\delta\omega}{\omega}.$$

But the nonrelativistic Doppler effect (the small β limit of (10.47)) reads

$$\omega' = \left(1 - \frac{v}{c}\cos\theta\right)\omega$$

or $(\delta\omega/\omega) = (v/c)\cos\theta$.

(9.1) **Another form of the expansion equation** Consider the energy balance equation (8.11)

$$\frac{1}{2}\dot{r}^2 - \frac{G_N M}{r} = \text{const.}$$

leading to

$$\dot{a}^2 - \frac{8\pi G_N}{3}\rho a^2 = \text{const}'.$$

which can also be obtained easily from (8.1). Dividing through by the second term and using the definition of critical density we have

$$\Omega^{-1} - 1 = \frac{\text{const.}}{\rho a^2}.$$

(9.2) **The epoch-dependent Hubble constant and $a(t)$** Using (8.7) to replace the curvature parameter k in the Friedmann Eq. (8.1), we have

$$\frac{\dot{a}^2(t)}{a^2(t)} = \frac{8\pi G_N}{3}\rho + \frac{\dot{a}^2(t_0)}{a^2(t)}(1 - \Omega_0) = H_0^2\left(\frac{\rho}{\rho_c} + \frac{1 - \Omega_0}{a^2(t)}\right). \tag{C.34}$$

Putting the time-dependence of the densities

$$\frac{\rho}{\rho_c} = \Omega(t) = \frac{\Omega_{R,0}}{a^4} + \frac{\Omega_{M,0}}{a^3} + \Omega_{\Lambda,0},$$

Eq. (C.34) becomes

$$\frac{H^2(t)}{H_0^2} = \frac{\Omega_{R,0}}{a^4} + \frac{\Omega_{M,0}}{a^3} + \Omega_{\Lambda,0} + \frac{1 - \Omega_0}{a^2}.$$

(9.4) **Negative Λ and the "big crunch"** For the $\Omega_0 = 1$ flat universe with matter and dark energy, we have the Friedmann Eq. (8.10)

$$H(a) = H_0[\Omega_{M,0}a^{-3} + \Omega_\Lambda]^{1/2}.$$

At $a = a_{\text{max}}$ the universe stops expanding and $H(a_{\text{max}}) = 0$, thus

$$a_{\text{max}} = \left(-\frac{\Omega_{\text{M},0}}{\Omega_\Lambda}\right)^{1/3}.$$

The cosmic time for the big crunch being twice the time for the universe to go from a_{max} to $a = 0$, we calculate in a way similar to (9.43)

$$2t_{\text{H}} \int_0^{a_{\text{max}}} \frac{da}{[\Omega_{\text{M},0}a^{-1} + \Omega_\Lambda a^2]^{1/2}} = \frac{4t_{\text{H}}}{3\sqrt{-\Omega_\Lambda}} \int_0^{a_{\text{max}}^{3/2}} \frac{dx}{[a_{\text{max}}^3 - x^2]^{1/2}}$$

$$= \frac{4t_{\text{H}}}{3\sqrt{-\Omega_\Lambda}} \left[\sin^{-1}\left(\frac{x}{a_{\text{max}}^{3/2}}\right)\right]_0^{a_{\text{max}}^{3/2}} = \frac{2\pi}{3\sqrt{-\Omega_\Lambda}} t_{\text{H}} = t_*.$$

(9.5) **Another estimate of deceleration/acceleration transition time** We define the matter and dark energy equality time $t_{\text{M}\Lambda}$ as

$$\rho_{\text{M}}(t_{\text{M}\Lambda}) = \rho_\Lambda(t_{\text{M}\Lambda}).$$

Using the scaling properties of these densities we have

$$\frac{\rho_{\text{M},0}}{a_{\text{M}\Lambda}^3} = \rho_{\Lambda,0}$$

or

$$1 + z_{\text{M}\Lambda} = (a_{\text{M}\Lambda})^{-1} = \left(\frac{\Omega_\Lambda}{\Omega_{\text{M},0}}\right)^{1/3}$$

which differs from the result in (9.47) only by a factor of $2^{1/3} \approx 1.2$.

(10.1) **Basis and inverse basis vectors: a simple exercise**

(a) Given the basis vectors

$$\mathbf{e}_1 = a\begin{pmatrix}1\\0\end{pmatrix}, \quad \mathbf{e}_2 = b\begin{pmatrix}\cos\theta\\ \sin\theta\end{pmatrix}.$$

The inverse basis vectors are

$$\mathbf{e}^1 = \frac{1}{a}(1, \quad -\cot\theta), \quad \mathbf{e}^2 = \frac{1}{b}(0, \quad \csc\theta).$$

The condition $\mathbf{e}_1 \cdot \mathbf{e}^1 = \mathbf{e}_2 \cdot \mathbf{e}^2 = 1$ and $\mathbf{e}_1 \cdot \mathbf{e}^2 = \mathbf{e}_2 \cdot \mathbf{e}^1 = 0$ can be easily checked by explicit vector multiplication. For example, $\mathbf{e}_2 \cdot \mathbf{e}^1 = (b/a)(\cos\theta - \cos\theta) = 0$

(b) Similarly by explicit vector multiplications, we have

$$g_{ij} = \mathbf{e}_i \cdot \mathbf{e}_j = \begin{pmatrix} a^2 & ab\cos\theta \\ ab\cos\theta & b^2 \end{pmatrix},$$

$$g^{ij} = \mathbf{e}^i \cdot \mathbf{e}^j = \begin{pmatrix} \dfrac{1}{a^2\sin^2\theta} & -\dfrac{\cos\theta}{ab\sin^2\theta} \\ -\dfrac{\cos\theta}{ab\sin^2\theta} & \dfrac{1}{b^2\sin^2\theta} \end{pmatrix}$$

so that $g_{ij}g^{jk} = \delta_{ik}$ can be checked by matrix multiplication.

(c) We can verify the completeness condition by calculating the direct-products of basis vectors,

$$\sum_i \mathbf{e}_i \otimes \mathbf{e}^i = \begin{pmatrix}1\\0\end{pmatrix}(1 \quad -\cot\theta) + \begin{pmatrix}\cos\theta\\ \sin\theta\end{pmatrix}(0 \quad \csc\theta)$$

$$= \begin{pmatrix}1 & -\cot\theta\\ 0 & 0\end{pmatrix} + \begin{pmatrix}0 & \cot\theta\\ 0 & 1\end{pmatrix} = \mathbf{1}.$$

(10.4) **Transformations: components vs. basis vectors** V^i transform "oppositely" from the bases vectors $\mathbf{e}_i$

$$\mathbf{e}_i \longrightarrow \mathbf{e}'_i = [\bar{\mathbf{L}}]_i^{\ j}\mathbf{e}_j \tag{C.35}$$

because the vector itself $\mathbf{V} = V^i\,\mathbf{e}_i$ does not change under the coordinate transformations.

(10.5) **g_{ij} is a tensor**

(a) Plugging in the transformations of the basis vectors (C.35) in the metric definition $g'_{ij} = \mathbf{e}'_i \cdot \mathbf{e}'_j$ we immediately obtain that for the metric, (10.13).

(b) The invariance of the scalar product $\mathbf{U}\cdot\mathbf{V}$ can also expressed as

$$U_iV_jg^{ij} = U'_kV'_lg'^{kl} = U_iV_j[\bar{\mathbf{L}}]^i_k[\bar{\mathbf{L}}]^j_lg'^{kl}$$

or

$$g^{ij} = [\bar{\mathbf{L}}]^i_k[\bar{\mathbf{L}}]^j_lg'^{kl}.$$

We can invert this equation by multiplying two [**L**] factors on both sides and using the relation Eq. (10.15) on the RHS to obtain

$$g'^{ij} = [\mathbf{L}]^i_k[\mathbf{L}]^j_lg^{kl}.$$

This shows, cf. Eq. (10.11), that the (inverse) metric is indeed a bona fide contravariant tensor.

(10.6) **The quotient theorem** Given that the product $U^iV^jg_{ij}$ is a scalar, and vectors U^i and V^j are known to be tensors, their quotient g_{ij} must also be a tensor.

(10.7) **Lorentz transform and velocity addition rule** Suppressing the transverse components, the 4-velocities have components $U^\mu = (U^0, U^1) = \gamma_u(c, u)$ and $U'^\mu = \gamma'_u(c, u')$, which are connected by Lorentz transformation $U'^\mu = [\mathbf{L}]^\mu_\nu U^\nu$:

$$\begin{pmatrix}\gamma'_uc\\ \gamma'_uu'\end{pmatrix} = \begin{pmatrix}\gamma_v & -\gamma_v\beta_v\\ -\gamma_v\beta_v & \gamma_v\end{pmatrix}\begin{pmatrix}\gamma_uc\\ \gamma_uu\end{pmatrix} = c\gamma_v\gamma_u\begin{pmatrix}1-\beta_v\beta_u\\ -\beta_v+\beta_u\end{pmatrix}.$$

Equating the first elements leads to $\gamma'_u = \gamma_v\gamma_u(1-\beta_v\beta_u)$. When this is substituted into the equality of the second elements, we obtain the velocity addition rule of (2.24).

(10.8) **Gravitational frequency shift: another derivation** The receiver being in motion, moving with nonrelativistic velocity $\beta = \Delta u/c$, the SR Lorentz frequency transformation (10.48) becomes

$$\frac{\omega_{\text{rec}}}{\omega_{\text{em}}} = \sqrt{\frac{1-\beta}{1+\beta}} \simeq 1-\beta.$$

Or, equivalently

$$\frac{\Delta\omega}{\omega} = -\frac{\Delta u}{c}$$

which is just the expression shown in (3.22).

(10.9) **Antiproton production threshold** The minimum energy needed to produce the final state of three protons and one antiproton in the center-of-mass frame is $4mc^2$. The square of the total 4-momentum of the final state, total momentum being zero, must then be $16m^2c^4$. By energy–momentum conservation, this must also be the square of the total 4-momentum of the initial state of two protons. If we denote the total energy and total 3-momentum of the initial state by $(E, \mathbf{p})$, we then have from (10.36)

$$16m^2c^4 = E^2 - |\mathbf{p}|^2c^2. \tag{C.36}$$

In the lab frame the target proton is at rest $\mathbf{p}_2 = 0$, we have

$$E = E_1 + mc^2, \quad |\mathbf{p}|^2c^2 = |\mathbf{p}_1|^2c^2 = E_1^2 - m^2c^4.$$

Substitute these two relations into (C.36) and solve for the projectile proton's lab energy $E_1 = 7mc^2$ to obtain its kinetic energy

$$K_{\text{lab}} = E_1 - mc^2 = 5.6\ \text{GeV}.$$

(10.10) **Covariant Lorentz force law**

$$K^0 = \frac{q}{c}F^{0i}U_i = \gamma\frac{q}{c}\mathbf{E}\cdot\mathbf{v} \tag{C.37}$$

is indeed $\gamma\mathbf{F}\cdot\mathbf{v}/c$ because the dot product with the magnetic field term in the Lorentz force vanishes.

(10.12) **Homogeneous Maxwell's equations** To show that $\partial_\mu F_{\nu\lambda}+\partial_\lambda F_{\mu\nu}+\partial_\nu F_{\lambda\mu} = 0$ follows from $\partial_\mu \tilde{F}^{\mu\nu} = 0$: from the definition of a dual field tensor, we have $\partial_\mu F_{\lambda\rho}\epsilon^{\mu\nu\lambda\rho} = 0$, which is a trivial relation $(0 = 0)$ if any pair of indices in (μ, λ, ρ) are equal. Thus, only when the indices are unequal do we get a nontrivial relation: take the example of equation of $\partial_\mu \tilde{F}^{\mu 0} = \partial_\mu F_{\lambda\rho}\epsilon^{\mu 0\lambda\rho} = 0$ we have

$$\partial_1 F_{23} + \partial_3 F_{12} + \partial_2 F_{31} = 0.$$

We can regard this as a relation in a particular coordinate frame with $\mu = 1$, $\nu = 2$, and $\lambda = 3$. Once written in the Lorentz covariant version, it must be valid in every frame. This is just the relation we set out to prove:

$$\partial_\mu F_{\nu\lambda} + \partial_\lambda F_{\mu\nu} + \partial_\nu F_{\lambda\mu} = 0.$$

To prove the converse statement, all we need to do is to contract $\epsilon^{\mu\nu\lambda\rho}$ onto the above equation.

(10.16) **$T^{\mu\nu}$ for a system of EM field and charges** We first calculate the divergence of $T^{\mu\nu}_{\text{charge}} = \rho'_{\text{mass}} U^\mu U^\nu$ to find that

$$\partial_\mu T^{\mu\nu}_{\text{charge}} = \rho'_{\text{mass}} (U^\mu \partial_\mu) U^\nu,$$

where we have also used the mass conservation law of $\partial_\mu (\rho'_{\text{mass}} U^\mu) = 0$. The Lorentz invariant product $U^\mu \partial_\mu$ can be evaluated in any convenient reference frame; we choose the comoving frame $U^\mu = \gamma(c, \mathbf{0})$ to obtain $U^\mu \partial_\mu = \gamma \partial_t = \partial_\tau$, the differentiation with respect to the proper time τ. The term $\rho'_{\text{mass}} \partial_\tau U^\nu$ is the 4-force density. Using the formula for the Lorentz force (density) of (10.60), we then have

$$\partial_\mu T^{\mu\nu}_{\text{charge}} = \rho'_{\text{mass}} \partial_\tau U^\nu = \frac{\rho'_{\text{charge}}}{c} F^{\nu\lambda} U_\lambda = \frac{1}{c} F^{\nu\lambda} j_\lambda,$$

where we have used the expression (10.75) for the electromagnetic current for free charges, $j_\lambda = \rho'_{\text{charge}} U_\lambda$.

We now calculate the divergence of $T^{\mu\nu}_{\text{field}}$ in (10.94) to find

$$\partial_\mu T^{\mu\nu}_{\text{field}} = \eta_{\alpha\beta} (\partial_\mu F^{\mu\alpha}) F^{\nu\beta}.$$

Here we have used the calculation performed in (10.96) and by noting the fact that, in the presence of charges, the inhomogeneous Maxwell's equation $\partial_\mu F^{\mu\alpha} = -(1/c) j^\alpha$ has a nonvanishing RHS:

$$\partial_\mu T^{\mu\nu}_{\text{field}} = -\frac{1}{c} \eta_{\alpha\beta} j^\alpha F^{\nu\beta} = -\frac{1}{c} F^{\nu\lambda} j_\lambda.$$

This shows clearly that the sum $T^{\mu\nu} = T^{\mu\nu}_{\text{field}} + T^{\mu\nu}_{\text{charge}}$ has zero divergence. Because fields and particles can exchange energy and momenta between them, energy and momentum are conserved only for the combined system.

(10.17) **Radiation pressure and energy density** The system of an electromagnetic field can be viewed either as a system of a field with energy–momentum tensor

$$T^{\mu\nu}_{\text{field}} = \eta_{\alpha\beta} F^{\mu\alpha} F^{\nu\beta} - \frac{1}{4} \eta^{\mu\nu} F^{\alpha\beta} F_{\alpha\beta}$$

or as a system of an ideal fluid made up of photons with, cf. (10.88),

$$T^{\mu\nu}_{\gamma\text{fluid}} = \begin{pmatrix} \rho' c^2 & & & \\ & p & & \\ & & p & \\ & & & p \end{pmatrix}$$

with $\rho' c^2$ and p being the radiation energy density and pressure, respectively. Since these two representations both describe the same system we should expect $T^{\mu\nu}_{\gamma\text{fluid}} = T^{\mu\nu}_{\text{field}}$, in particular their traces should be equal: $\eta_{\mu\nu} T^{\mu\nu}_{\gamma\text{fluid}} = \eta_{\mu\nu} T^{\mu\nu}_{\text{field}}$. But a simple inspection shows that $\eta_{\mu\nu} T^{\mu\nu}_{\text{field}} = 0$ because $\eta_{\mu\nu} \eta^{\mu\nu} = 4$. The vanishing trace $\eta_{\mu\nu} T^{\mu\nu}_{\gamma\text{fluid}} = 0$ leads to the result $p = \rho' c^2 / 3$.

(11.1) **Covariant derivative for a covariant vector** Given that $V_\mu V^\mu$ is an invariant, in the notation of (11.43), we also have $[\Delta(V_\mu V^\mu)]_{\text{coord}} = 0$:

$$V_\mu[\Delta V^\mu]_{\text{coord}} + V^\mu[\Delta V_\mu]_{\text{coord}} = 0.$$

Substituting (11.44) in $[\Delta V^\mu]_{\text{coord}} = -\Gamma^\mu_{\nu\lambda}V^\nu dx^\lambda$, we get

$$V^\mu[\Delta V_\mu]_{\text{coord}} = V_\mu \Gamma^\mu_{\nu\lambda} V^\nu\, dx^\lambda = V^\mu(\Gamma^\nu_{\mu\lambda}V_\nu\, dx^\lambda).$$

The last expression is reached by relabelling $\mu \leftrightarrow \nu$. The result $[\Delta V_\mu]_{\text{coord}} = +\Gamma^\nu_{\mu\lambda}V_\nu dx^\lambda$ implies that

$$D_\nu V_\mu = \partial_\nu V_\mu - \Gamma^\lambda_{\nu\mu}V_\lambda.$$

(11.2) **Moving bases and Christoffel symbols in polar coordinates for a flat plane**

(a) Explicitly differentiating the relation

$$\mathbf{r} = r\cos\theta\mathbf{i} + r\sin\theta\mathbf{j}$$

we have

$$d\mathbf{r} \equiv dr\,\mathbf{e}_r + d\theta\mathbf{e}_\theta = dr\cos\theta\mathbf{i} - r\sin\theta d\theta\,\mathbf{i} + dr\sin\theta\mathbf{j} + r\cos\theta d\theta\,\mathbf{j}.$$

Collecting the dr and $d\theta$ terms,

$$\mathbf{e}_r = \cos\theta\mathbf{i} + \sin\theta\,\mathbf{j},$$
$$\mathbf{e}_\theta = -r\sin\theta\mathbf{i} + r\cos\theta\mathbf{j}.$$

The inverse bases can be found by contracting with the inverse metric $g^{\mu\nu} = \text{diag}(1,\ r^{-2})$:

$$\mathbf{e}^r = \cos\theta\mathbf{i} + \sin\theta\mathbf{j},$$
$$\mathbf{e}^\theta = -r^{-1}\sin\theta\mathbf{i} + r^{-1}\cos\theta\mathbf{j}.$$

(b) To calculate the Christoffel symbols through their definition of $\partial_\nu\mathbf{e}^\mu = -\Gamma^\mu_{\nu\lambda}\mathbf{e}^\lambda$ we first observe:

$$\frac{\partial\mathbf{e}^r}{\partial r} = 0, \quad \frac{\partial\mathbf{e}^\theta}{\partial r} = \frac{-1}{r^2}(-\sin\theta\mathbf{i} + \cos\theta\mathbf{j}) = \frac{-1}{r}\mathbf{e}^\theta.$$

Then the definitions

$$\frac{\partial\mathbf{e}^r}{\partial r} = \Gamma^r_{rr}\mathbf{e}^r + \Gamma^r_{r\theta}\mathbf{e}^\theta, \quad \frac{\partial\mathbf{e}^\theta}{\partial r} = \Gamma^\theta_{rr}\mathbf{e}^r + \Gamma^\theta_{r\theta}\mathbf{e}^\theta$$

allow us to read off the Christoffel symbols $\Gamma^r_{rr} = \Gamma^r_{r\theta} = \Gamma^\theta_{rr} = 0$ and $\Gamma^\theta_{r\theta} = 1/r$. Similarly, from

$$\frac{\partial\mathbf{e}^r}{\partial\theta} = -\sin\theta\mathbf{i} + \cos\theta\,\mathbf{j} = r\mathbf{e}^\theta,$$
$$\frac{\partial\mathbf{e}^\theta}{\partial\theta} = -r^{-1}\cos\theta\mathbf{i} - r^{-1}\sin\theta\,\mathbf{j} = -r^{-1}\mathbf{e}^r$$

we obtain $\Gamma^r_{\theta r} = \Gamma^\theta_{\theta\theta} = 0$, $\Gamma^r_{\theta\theta} = -r$ and $\Gamma^\theta_{\theta r} = r^{-1}$.

(c) Work out the components in

$$\begin{aligned}D_\mu V^\mu &= \partial_\mu V^\mu + \Gamma^\mu_{\mu\nu} V^\nu \\ &= \partial_r V^r + \partial_\theta V^\theta + (\Gamma^r_{rr} + \Gamma^\theta_{\theta r})V^r + (\Gamma^r_{r\theta} + \Gamma^\theta_{\theta\theta})V^\theta \\ &= \partial_r V^r + \partial_\theta V^\theta + \frac{1}{r}V^r = \frac{1}{r}\frac{\partial}{\partial r}(rV^r) + \frac{\partial}{\partial\theta}V^\theta \\ &= \begin{pmatrix}\frac{1}{r}\frac{\partial}{\partial r}r & \frac{\partial}{\partial\theta}\end{pmatrix}\begin{pmatrix}V^r \\ V^\theta\end{pmatrix}.\end{aligned}$$

(d) Because the scalar function $\Phi(x)$ is coordinate independent, $D_\mu\Phi = \partial_\mu\Phi$. To raise the index we must multiply it by the inverse metric $g^{\mu\nu}\partial_\mu\Phi$. Using the result obtained in (c) we have

$$\begin{aligned}D_\mu D^\mu \Phi(x) &= D_\mu(g^{\mu\nu}\partial_\mu\Phi) \\ &= \begin{pmatrix}\frac{1}{r}\frac{\partial}{\partial r}r & \frac{\partial}{\partial\theta}\end{pmatrix}\begin{pmatrix}1 & 0 \\ 0 & r^{-2}\end{pmatrix}\begin{pmatrix}\partial_r\Phi \\ \partial_\theta\Phi\end{pmatrix} \\ &= \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial\Phi}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2\Phi}{\partial\theta^2}.\end{aligned}$$

(e) The metric in polar coordinates has only one nontrivial element $g_{\theta\theta} = r^2$. Checking the covariant differentiation with respect to the radial coordinate r, we get

$$D_r g_{\theta\theta} = \partial_r g_{\theta\theta} - 2\Gamma^\mu_{r\theta} g_{\mu\theta} = 2r - 2\frac{1}{r}r^2 = 0.$$

(f) Substituting $g_{\theta r} = 0$ and $g_{\theta\theta} = r^2$ into (11.37), we have

$$\begin{aligned}\Gamma^r_{\theta\theta} &= \frac{1}{2}g^{r\mu}(\partial_\theta g_{\theta\mu} + \partial_\theta g_{\theta\mu} - \partial_\mu g_{\theta\theta}) \\ &= \frac{1}{2}g^{rr}(2\partial_\theta g_{\theta r} - \partial_r g_{\theta\theta}) = -r, \\ \Gamma^\theta_{\theta\theta} &= \frac{1}{2}g^{\theta\theta}\partial_\theta g_{\theta\theta} = 0.\end{aligned}$$

(11.3) **Symmetry property of Christoffel symbols** Because a scalar field $\Phi(x)$ is coordinate-independent, there is no difference between their covariant and ordinary derivatives, $D_\mu\Phi = \partial_\mu\Phi$. We then apply the result of Problem 11.1 to obtain

$$D_\nu D_\mu \Phi = \partial_\nu\partial_\mu\Phi - \Gamma^\lambda_{\nu\mu}\partial_\lambda\Phi.$$

Because the first two terms are manifestly symmetric in (μ, ν), the last term (i.e. $\Gamma^\lambda_{\nu\mu}$) must also be symmetric in (μ, ν).

(11.4) **Metric is covariantly constant: further proofs**

(a) Take the covariant derivative of the metric tensor (with covariant indices) and then express the resulting Christoffel symbols in terms of derivatives of the metric:

$$\begin{aligned}D_\mu g_{\nu\lambda} &= \partial_\mu g_{\nu\lambda} - \Gamma^\rho_{\mu\nu} g_{\rho\lambda} - \Gamma^\rho_{\mu\lambda} g_{\rho\nu}\\ &= \partial_\mu g_{\nu\lambda} - \frac{1}{2} g^{\rho\sigma}(\partial_\mu g_{\nu\sigma} + \partial_\nu g_{\mu\sigma} - \partial_\sigma g_{\mu\nu}) g_{\rho\lambda}\\ &\quad - \frac{1}{2} g^{\rho\sigma}(\partial_\mu g_{\lambda\sigma} + \partial_\lambda g_{\mu\sigma} - \partial_\sigma g_{\mu\lambda}) g_{\rho\nu}.\end{aligned}$$

After summing over repeated indices, we find all terms cancel.

(b) The metric's first derivatives and the connection symbol vanish in the locally Euclidean coordinates: $\partial_\lambda g_{\mu\nu} = 0$ and $\Gamma^\mu_{\nu\lambda} = 0$. We thus have $D_\lambda g_{\mu\nu} = 0$ in the LEF frame. Since this is a covariant equation, it must be valid in **every frame**.

(11.5) **$D_\nu V_\mu$ is a good tensor: another proof** We can use the geodesic equation in the form of $(D/D\sigma) \times (dx^\mu/d\sigma) = 0$ to obtain

$$\frac{D}{D\sigma}\left(V_\mu \frac{dx^\mu}{d\sigma}\right) = \left(\frac{DV_\mu}{D\sigma}\right)\frac{dx^\mu}{d\sigma} = 0.$$

which may be written as

$$(D_\nu V_\mu)\frac{dx^\mu}{d\sigma}\frac{dx^\nu}{d\sigma} = 0.$$

The quotient theorem then informs us that $D_\nu V_\mu$ is a good tensor, because it is contracted into a good tensor: $(dx^\mu/d\sigma)(dx^\nu/d\sigma)$.

(11.6) **Parallel transport of a vector around a general spherical triangle** The triangle has three vertices (A, B, C) connected by geodesic curves with interior angles (α, β, γ). We now transport a vector around this triangle, along the three geodesic sides of the triangle. The key observation is that the angle subtended by the vector and the geodesic is unchanged (cf. the worked example in the text).

1. At vertex A, the vector makes an angle θ_1 with the tangent along AB.
2. At vertex B, the vector makes the same angle θ_1 with the tangent along AB, thus it makes $\theta_2 = \theta_1 + (\pi - \beta)$ along BC.
3. At vertex C, the vector makes $\theta_3 = \theta_2 + (\pi - \gamma)$ along CA.
4. Returning to A, the vector makes $\theta_4 = \theta_3 + (\pi - \alpha)$ along the original AB.

Plug in θ_i sequentially and take out a trivial factor of 2π, we obtain the directional change of the vector

$$\delta\theta = \theta_1 - \theta_4 = \alpha + \beta + \gamma - \pi,$$

which is just the angular excess ϵ.

(11.9) **Counting independent elements of Riemann tensor** Write the curvature tensor as $R_{\{[\mu\nu],[\alpha\beta]\}}$ to remind ourselves of the symmetry properties of (11.69)–(11.71): antisymmetry of Eq. (11.69) as $[\mu\nu]$, that of (11.70) as $[\alpha\beta]$, and the symmetry of (11.71) as $\{[\mu\nu],[\alpha\beta]\}$. An $n \times n$ matrix has $\frac{1}{2}n(n+1)$ independent elements if it is symmetric, and $\frac{1}{2}n(n-1)$ elements if antisymmetric. Hence, for the purpose of counting independent components, we can regard $R_{\{[\mu\nu],[\alpha\beta]\}}$ as a $\frac{1}{2}n(n-1)$ by $\frac{1}{2}n(n-1)$ matrix, which is symmetric. This yields a count of

$$M_{(n)} = \frac{1}{2}\left[\frac{1}{2}n(n-1)\right] \times \left[\frac{1}{2}n(n-1)+1\right]$$
$$= \frac{1}{8}n(n-1)(n^2-n+2).$$

There are not as many independent elements as $M_{(n)}$ because we also need to factor-in further the cyclic symmetry constraint of (11.72). Actually, (11.72) represents extra conditions that reduce the number of independent elements **only if** all four indices are different—because otherwise this cyclic condition reduces to the first three symmetry conditions. Thus the number of additional constraint conditions as represented by (11.72) is given by:

$$C_{(n)} = \binom{n}{4} = \frac{n(n-1)(n-2)(n-3)}{4!}.$$

Subtracting $C_{(n)}$ from $M_{(n)}$ leads to the the number of independent components of a curvature tensor in an n-dimensional space:

$$N_{(n)} = M_{(n)} - C_{(n)} = \frac{1}{12}n^2(n^2-1). \qquad \text{(C.38)}$$

For the 4D spacetime, $N_{(4)} = 20$.

(11.10) **The number of metric's independent second derivatives and Riemann tensor**

(a) Remembering that the number of independent elements of a symmetric $n \times n$ matrix is $n(n+1)/2$, we see that the tensor $g_{\mu\nu}$ has 10 elements, and its first derivative $\partial_\alpha g_{\mu\nu}$ has 40, and its second derivative $\partial_\alpha \partial_\beta g_{\mu\nu}$ has 100 elements, when we used the fact that $\partial_\alpha \partial_\beta = \partial_\beta \partial_\alpha$. Namely,

	Index sym	$A_{(4)}$
$g_{\mu\nu}$	$\{\mu\nu\}$	$(4 \times 5)/2 = 10$
$\partial_\alpha g_{\mu\nu}$	$\alpha\{\mu\nu\}$	$4 \times 10 = 40$
$\partial_\alpha \partial_\beta g_{\mu\nu}$	$\{\alpha\beta\}\{\mu\nu\}$	$10 \times 10 = 100$

In particular the number of components for the second derivative $\partial_\alpha \partial_\beta g_{\mu\nu}$ in an n-dimensional space is

$$A_{(n)} = \left[\frac{1}{2}n(n+1)\right]^2. \qquad \text{(C.39)}$$

(b) Using the same notation as in (a), we find the number of parameters in the transformations for the 4D space:

	Index sym	$B_{(4)}$
$(\partial_\alpha x_\beta)$	$\alpha\beta$	$4 \times 4 = 16$
$\partial_\gamma(\partial_\alpha x_\beta)$	$\{\gamma\alpha\}\beta$	$10 \times 4 = 40$
$\partial_\gamma \partial_\delta(\partial_\alpha x_\beta)$	$\{\alpha\gamma\delta\}\beta$	$20 \times 4 = 80$

where, on the last line for the second derivative $\partial_\gamma \partial_\delta (\partial_\alpha x_\beta)$, we have used the fact that there are 20 possible totally symmetric combinations of three indices ($d = 3$) when each index can take on four possible values ($n = 4$). This is an example of the general result $N(d, n)$ being the number of symmetric combinations of d objects each can take on n possible values:

$$N(d,n) = \binom{d+n-1}{d} = \frac{(n+d-1)!}{d!(n-1)!}. \qquad \text{(C.40)}$$

One can understand this result by thinking of the ways, for example, of placing d identical balls into n boxes, which is equivalent to the problem of permuting d identical balls and $n - 1$ identical partitions.

(c) After a comparison of the results obtained in (a) and (b)

	$\mathrm{A}_{(4)}$		$\mathrm{B}_{(4)}$
$g_{\mu\nu}$	10	$(\partial_\alpha x_\beta)$	16
$\partial_\alpha g_{\mu\nu}$	40	$\partial_\gamma (\partial_\alpha x_\beta)$	40
$\partial_\alpha \partial_\beta g_{\mu\nu}$	100	$\partial_\gamma \partial_\delta (\partial_\alpha x_\beta)$	80

we comment on each case:

i. *The $g_{\mu\nu}$ case*: Do we need the 16 parameters of $(\partial_\alpha x_\beta)$ to determine the 10 elements of $g_{\mu\nu}$? Yes, because the transformation includes the six parameter Lorentz transformations that leave the Euclidean metric $g_{\mu\nu} = \eta_{\mu\nu}$ invariant.

ii. *The $\partial_\alpha g_{\mu\nu}$ case*: There are just the correct number (40) of parameters in $\partial_\gamma (\partial_\alpha x_\beta)$ to set all the 40 independent elements of $\partial_\alpha g_{\mu\nu}$ to zero. (See the flatness theorem.)

iii. *The $\partial_\alpha \partial_\beta g_{\mu\nu}$ case*: We still have 20 yet undetermined elements in the second derivative $\partial_\alpha \partial_\beta g_{\mu\nu}$. This corresponds to the number of independent elements in the 4D curvature tensor $N_{(4)} = 20$ as shown in Problem 11.9.

(d) For a general n-dimensional space, the number of second derivatives of the transformation $\partial_\gamma \partial_\delta \partial_\alpha x_\beta$ as given by (C.40) for $d = 3$ (with a further multiplication of n for the β index) is

$$\mathrm{B}_{(n)} = \frac{1}{6} n^2 (n+2)(n+1). \qquad \text{(C.41)}$$

The number of independent elements of the second derivative must be the difference of (C.39) and (C.41)

$$N_{(n)} = \mathrm{A}_{(n)} - \mathrm{B}_{(n)} = \frac{1}{12} n^2 (n^2 - 1),$$

which exactly matches the result of (C.38).

(11.11) **Reducing Riemann tensor to Gaussian curvature** For a 2D space with orthogonal coordinates, we have the metrics

$$g_{\mu\nu} = \begin{pmatrix} g_{11} & 0 \\ 0 & g_{22} \end{pmatrix}, \quad g^{\mu\nu} = \begin{pmatrix} g^{11} & 0 \\ 0 & g^{22} \end{pmatrix}$$

with $g^{11} = 1/g_{11}$ and $g^{22} = 1/g_{22}$ so that $g_{\mu\nu}g^{\nu\lambda} = \delta^{\lambda}_{\mu}$. The Christoffel symbols can be calculated from

$$\Gamma^1_{\mu\nu} = \frac{1}{2}g^{11}(\partial_\mu g_{1\nu} + \partial_\nu g_{1\mu} - \partial_1 g_{\mu\nu})$$

so that

$$\begin{aligned}\Gamma^1_{11} &= \frac{1}{2g_{11}}\partial_1 g_{11}, \quad \Gamma^1_{22} = -\frac{1}{2g_{11}}\partial_1 g_{22},\\ \Gamma^1_{12} &= \Gamma^1_{21} = \frac{1}{2g_{11}}\partial_2 g_{11}.\end{aligned}$$

Similarly, we also have

$$\Gamma^2_{22} = \frac{1}{2g_{221}}\partial_2 g_{22}, \quad \Gamma^2_{12} = \Gamma^2_{21} = \frac{1}{2g_{22}}\partial_1 g_{22}.$$

The only nontrivial (and independent) curvature element is

$$\begin{aligned}R_{1212} &= g_{1\mu}R^{\mu}_{212}\\ &= g_{11}(\partial_2\Gamma^1_{21} - \partial_1\Gamma^1_{22} + \Gamma^\nu_{21}\Gamma^1_{\nu 2} - \Gamma^\nu_{22}\Gamma^1_{\nu 1})\\ &= g_{11}(\partial_2\Gamma^1_{21} - \partial_1\Gamma^1_{22} + \Gamma^1_{21}\Gamma^1_{12} + \Gamma^2_{21}\Gamma^1_{22}\\ &\quad - \Gamma^1_{22}\Gamma^1_{1\nu 1} - \Gamma^2_{22}\Gamma^1_{21})\\ &= \frac{1}{2}\left\{\partial_2^2 g_{11} + \partial_1^2 g_{22} - \frac{1}{2g_{11}}[(\partial_1 g_{11})(\partial_1 g_{22}) + (\partial_2 g_{11})^2]\right.\\ &\quad \left. - \frac{1}{2g_{22}}[(\partial_2 g_{11})(\partial_2 g_{22}) + (\partial_1 g_{22})^2]\right\}\end{aligned}$$

which, when divided by the metric determinant $\det g = g_{11}g_{22}$, the ratio $-R_{1212}/\det g$ is recognized as the Gaussian curvature of (4.35).

(11.14) **Contraction of Christoffel symbols** The inverse matrix $[g_{\mu\nu}]^{-1}$ has elements $g^{\mu\nu}$, which are related to the determinant g of the matrix $[g_{\mu\nu}]$ and the cofactors $C^{\mu\nu}$ (associated with elements $g_{\mu\nu}$) as

$$g^{\mu\nu} = \frac{C^{\mu\nu}}{g}. \tag{C.42}$$

Also, the determinant g can be expanded as (for any fixed μ)

$$g = \sum_\nu g_{\mu\nu}C^{\mu\nu}, \tag{C.43}$$

where we have displayed the summation sign to emphasize that there is no summation over the index μ. Because the determinant is a function of the matrix elements $g_{\mu\nu}$ which in turn are position dependent, we have

$$\frac{\partial g}{\partial x^\alpha} = \frac{\partial g}{g_{\mu\nu}}\frac{\partial g_{\mu\nu}}{\partial x^\alpha} = C^{\mu\nu}\frac{\partial g_{\mu\nu}}{\partial x^\alpha} = gg^{\mu\nu}\partial_\alpha g_{\mu\nu}, \tag{C.44}$$

where we have used (C.43) and (C.42) to reach the last two expressions. Knowing this identity, we proceed to make a contraction of the Christoffel symbols

$$\Gamma^{\mu}_{\mu\alpha} = \frac{1}{2}g^{\mu\nu}[\partial_\alpha g_{\mu\nu} + \partial_\mu g_{\alpha\nu} - \partial_\nu g_{\mu\alpha}].$$

The last two terms $\partial^\nu g_{\alpha\nu} = \partial^\mu g_{\mu\alpha}$ cancel so that the contraction can be rewritten by (C.44) as

$$\Gamma^{\mu}_{\mu\alpha} = \frac{1}{2}g^{\mu\nu}\partial_\alpha g_{\mu\nu} = \frac{1}{2g}\frac{\partial g}{\partial x^\alpha}$$

which is equivalent to the sought after result of

$$\Gamma^{\mu}_{\mu\alpha} = \frac{1}{\sqrt{-g}}\frac{\partial}{\partial x^\alpha}\sqrt{-g}.$$

(11.15) **Contraction of Riemann tensor** Contracting the first two indices $R^{\mu}_{\ \mu\alpha\beta}$ (11.58):

$$\partial_\alpha\Gamma^{\mu}_{\mu\beta} - \partial_\beta\Gamma^{\mu}_{\mu\alpha} + \Gamma^{\mu}_{\nu\alpha}\Gamma^{\nu}_{\mu\beta} - \Gamma^{\mu}_{\nu\beta}\Gamma^{\nu}_{\mu\alpha}.$$

The dummy indices in the last two terms can be relabeled $\mu \leftrightarrow \nu$; we see that they cancel each other. A straightforward calculation of the first two terms by using the result obtained in Problem 11.14 shows that they cancel each other also.

(12.4) **The equation of geodesic deviation** Let us consider two particles: one has the spacetime trajectory x^μ and another has $x^\mu + s^\mu$. These two particles, separated by the displacement vector s^μ, obey the respective equations of motion:

$$\frac{d^2x^\mu}{d\tau^2} + \Gamma^{\mu}_{\alpha\beta}(x)\frac{dx^\alpha}{d\tau}\frac{dx^\beta}{d\tau} = 0$$

and

$$\left(\frac{d^2x^\mu}{d\tau^2} + \frac{d^2s^\mu}{d\tau^2}\right) + \Gamma^{\mu}_{\alpha\beta}(x+s)\left(\frac{dx^\alpha}{d\tau} + \frac{ds^\alpha}{d\tau}\right)\left(\frac{dx^\beta}{d\tau} + \frac{ds^\beta}{d\tau}\right) = 0.$$

When the separation distance s^μ is small, we can approximate the Christoffel symbols $\Gamma^{\mu}_{\alpha\beta}(x+s)$ by a Taylor expansion

$$\Gamma^{\mu}_{\alpha\beta}(x+s) = \Gamma^{\mu}_{\alpha\beta}(x) + \partial_\lambda\Gamma^{\mu}_{\alpha\beta}s^\lambda + \cdots .$$

From the difference of the two geodesic equations, we obtain, to first order in s^μ,

$$\frac{d^2s^\mu}{d\tau^2} = -2\Gamma^{\mu}_{\alpha\beta}\frac{ds^\alpha}{d\tau}\frac{dx^\beta}{d\tau} - \partial_\lambda\Gamma^{\mu}_{\alpha\beta}s^\lambda\frac{dx^\alpha}{d\tau}\frac{dx^\beta}{d\tau}. \tag{C.45}$$

The relative acceleration is the second derivative of the separation s^μ along the worldline (i.e. the double differentiation along the geodesic curve). From (11.46) we have the first derivative

$$\frac{Ds^\mu}{D\tau} = \frac{ds^\mu}{d\tau} + \Gamma^\mu_{\alpha\beta} s^\alpha \frac{dx^\beta}{d\tau}$$

and the second derivative

$$\begin{aligned}
\frac{D^2 s^\mu}{D\tau^2} &= \frac{D}{D\tau}\left(\frac{Ds^\mu}{D\tau}\right) = \frac{d}{d\tau}\left(\frac{Ds^\mu}{D\tau}\right) + \Gamma^\mu_{\alpha\beta}\left(\frac{Ds^\alpha}{D\tau}\right)\frac{dx^\beta}{d\tau} \\
&= \frac{d}{d\tau}\left(\frac{ds^\mu}{d\tau} + \Gamma^\mu_{\alpha\beta} s^\alpha \frac{dx^\beta}{d\tau}\right) + \Gamma^\mu_{\alpha\beta}\left(\frac{ds^\alpha}{d\tau} + \Gamma^\alpha_{\lambda\rho} s^\lambda \frac{dx^\rho}{d\tau}\right)\frac{dx^\beta}{d\tau} \\
&= \frac{d^2 s^\mu}{d\tau^2} + \partial_\lambda \Gamma^\mu_{\alpha\beta}\frac{dx^\lambda}{d\tau} s^\alpha \frac{dx^\beta}{d\tau} + \Gamma^\mu_{\alpha\beta}\frac{ds^\alpha}{d\tau}\frac{dx^\beta}{d\tau} + \Gamma^\mu_{\alpha\beta} s^\alpha \frac{d^2 x^\beta}{d\tau^2} \\
&\quad + \Gamma^\mu_{\alpha\beta}\frac{ds^\alpha}{d\tau}\frac{dx^\beta}{d\tau} + \Gamma^\mu_{\alpha\beta}\Gamma^\alpha_{\lambda\rho} s^\lambda \frac{dx^\rho}{d\tau}\frac{dx^\beta}{d\tau}. \qquad \text{(C.46)}
\end{aligned}$$

For the $d^2 s^\mu / d\tau^2$ term we use (C.45); for the $d^2 x^\beta / d\tau^2$ term we use the geodesic equation

$$\frac{d^2 x^\beta}{d\tau^2} = -\Gamma^\beta_{\lambda\rho}\frac{dx^\lambda}{d\tau}\frac{dx^\rho}{d\tau}.$$

This way one finds

$$\begin{aligned}
\frac{D^2 s^\mu}{D\tau^2} &= -2\Gamma^\mu_{\alpha\beta}\frac{ds^\alpha}{d\tau}\frac{dx^\beta}{d\tau} - \partial_\lambda \Gamma^\mu_{\alpha\beta} s^\lambda \frac{dx^\alpha}{d\tau}\frac{dx^\beta}{d\tau} + \partial_\lambda \Gamma^\mu_{\alpha\beta}\frac{dx^\lambda}{d\tau} s^\alpha \frac{dx^\beta}{d\tau} \\
&\quad + 2\Gamma^\mu_{\alpha\beta}\frac{ds^\alpha}{d\tau}\frac{dx^\beta}{d\tau} - \Gamma^\mu_{\alpha\beta} s^\alpha \Gamma^\beta_{\lambda\rho}\frac{dx^\lambda}{d\tau}\frac{dx^\rho}{d\tau} \\
&\quad + \Gamma^\mu_{\alpha\beta}\Gamma^\alpha_{\lambda\rho} s^\lambda \frac{dx^\rho}{d\tau}\frac{dx^\beta}{d\tau}.
\end{aligned}$$

After a cancellation of two terms and relabeling of several dummy indices, this becomes

$$\begin{aligned}
\frac{D^2 s^\mu}{D\tau^2} &= -\partial_\lambda \Gamma^\mu_{\alpha\beta} s^\lambda \frac{dx^\alpha}{d\tau}\frac{dx^\beta}{d\tau} + \partial_\alpha \Gamma^\mu_{\lambda\beta}\frac{dx^\alpha}{d\tau} s^\lambda \frac{dx^\beta}{d\tau} \\
&\quad - \Gamma^\mu_{\lambda\rho} s^\lambda \Gamma^\rho_{\alpha\beta}\frac{dx^\alpha}{d\tau}\frac{dx^\beta}{d\tau} + \Gamma^\mu_{\rho\beta}\Gamma^\rho_{\lambda\alpha} s^\lambda \frac{dx^\alpha}{d\tau}\frac{dx^\beta}{d\tau}
\end{aligned}$$

or

$$\frac{D^2 s^\mu}{D\tau^2} = -R^\mu_{\ \alpha\lambda\beta} s^\lambda \frac{dx^\alpha}{d\tau}\frac{dx^\beta}{d\tau},$$

where

$$R^\mu_{\alpha\lambda\beta} = \partial_\lambda \Gamma^\mu_{\alpha\beta} - \partial_\beta \Gamma^\mu_{\lambda\alpha} + \Gamma^\mu_{\lambda\rho}\Gamma^\rho_{\alpha\beta} - \Gamma^\mu_{\beta\rho}\Gamma^\rho_{\lambda\alpha}$$

in agreement with (11.58).

(12.5) **From geodesic deviation to NR tidal forces** Besides slow moving particles, the Newtonian limit means a weak gravitational field: $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$ with $h_{\mu\nu}$ being small. Thus (11.37) becomes

$$\Gamma^\mu_{\alpha\beta} = \frac{1}{2}\eta^{\mu\rho}[\partial_\alpha h_{\beta\rho} + \partial_\beta h_{\alpha\rho} - \partial_\rho h_{\alpha\beta}].$$

Also, in this weak-field limit, we can drop the quadratic terms in the curvature so that there are only two terms, related by the interchange of (β,λ) indices

$$R^{\mu}_{\alpha\lambda\beta} = \partial_\lambda \Gamma^{\mu}_{\alpha\beta} - \partial_\beta \Gamma^{\mu}_{\lambda\alpha}$$
$$= \frac{1}{2}\eta^{\mu\rho}[\partial_\lambda\partial_\alpha h_{\beta\rho} - \partial_\lambda\partial_\rho h_{\alpha\beta} - \partial_\beta\partial_\alpha h_{\lambda\rho} + \partial_\beta\partial_\rho h_{\alpha\lambda}]$$

after cancelling two terms. Thus

$$R^{i}_{0j0} = \frac{1}{2}[\partial_j\partial_0 h_{0i} - \partial_j\partial_i h_{00} - \partial_0\partial_0 h_{ji} + \partial_0\partial_i h_{0j}] = -\frac{1}{2}\partial_i\partial_j h_{00}.$$

Because the Newtonian limit also has the static field condition, to reach the last line we have dropped all time derivatives. With $h_{00} = -2\Phi/c^2$ as given by (5.20), we have the sought-after relation of

$$R^{j}_{0j0} = \frac{1}{c^2}\frac{\partial^2\Phi}{\partial x^i \partial x^j}.$$

(13.1) **Gauge transformations**

(a) Consider a coordinate (gauge) transformation as given in (13.12) so that, according to (13.17), $h'_{\alpha\beta} = h_{\alpha\beta} - \partial_\alpha\chi_\beta - \partial_\beta\chi_\alpha$. This implies (by contracting the indices on both sides) the transformation for the trace $h' = h - 2\partial^\beta\chi_\beta$. These two relations can be combined to yield the gauge transformation of $\bar{h}_{\alpha\beta}$,

$$h'_{\alpha\beta} - \frac{h'}{2}\eta_{\alpha\beta} = \bar{h}'_{\alpha\beta} = \bar{h}_{\alpha\beta} - \partial_\alpha\chi_\beta - \partial_\beta\chi_\alpha + \eta_{\alpha\beta}(\partial\chi). \tag{C.47}$$

(b) Taking the derivative on both sides of (C.47),

$$\partial^\alpha \bar{h}'_{\alpha\beta} = \partial^\alpha \bar{h}_{\alpha\beta} - \Box\chi_\beta$$

the new metric perturbation field can be made to obey the Lorentz condition $\partial^\alpha \bar{h}'_{\alpha\beta} = 0$ if

$$\Box\chi_\beta = \partial^\alpha \bar{h}_{\alpha\beta}.$$

(c) Plugging $\bar{h}_{\mu\nu} = \epsilon_{\mu\nu}e^{ikx}$ and $\chi_\nu = X_\nu e^{ikx}$ into the gauge transformation (C.47), we have

$$\epsilon'_{\mu\nu} = \epsilon_{\mu\nu} - ik_\mu X_\nu - ik_\nu X_\mu + i\eta_{\mu\nu}(k\cdot X) \tag{C.48}$$

which implies the trace relation

$$\epsilon'^{\,\mu}_{\mu} = \epsilon^{\mu}_{\mu} - 2ik^\mu X_\mu.$$

This means that if we start with a polarization tensor that is not traceless, it will be traceless $\epsilon'^{\mu}_{\mu} = 0$ in a new coordinate if the gauge vector function X_μ for the coordinate transformation is chosen to satisfy the condition $2ik^\mu X_\mu = \epsilon^{\mu}_{\mu}$. Now we have used one of the four numbers in X_μ to fix the trace. How can we use the remaining three to obtain $\epsilon_{\mu 0} = 0$ which would seem to represent four conditions? This is possible because we are working in the Lorentz gauge and k^μ is a null-vector. Here is the reason.

Starting with $\epsilon_{\mu 0} \neq 0$, new coordinate transformation leads to (C.48) with

$$\epsilon'_{\mu 0} = \epsilon_{\mu 0} - ik_\mu X_0 - ik_0 X_\mu + i\eta_{\mu 0}(k \cdot X).$$

Formally $\epsilon'_{\mu 0} = 0$ represents four conditions. But, because of $k^\mu \epsilon_{\mu 0} = 0$ and $k^2 = 0$, these four equations must obey a constraint relation, obtained by a contraction with the vector k^μ:

$$k^\mu \epsilon_{\mu 0} - ik^2 X_0 - ik_0(k \cdot X) + ik_0(k \cdot X) = 0.$$

Thus $\epsilon'_{\mu 0} = 0$ actually stands for three independent relations.

(d) The polarization tensor being symmetric, $\epsilon_{\mu\nu} = \epsilon_{\nu\mu}$, it has 10 independent elements. The Lorentz gauge condition $k^\mu \epsilon_{\mu\nu} = 0$ represents 4 constraints, $\epsilon_\mu^{\ \mu} = 0$ is one, and $\epsilon_{\mu 0} = 0$, as discussed above, is three. Thus there are only $10 - 4 - 1 - 3 = 2$ independent elements in the polarization tensor.

(13.2) **Wave effect via the deviation equation** With a collection of nearby particles, we can consider velocity and separation fields, $U^\mu(x)$ and $S^\mu(x)$. The equation of geodesic deviation (Problem 12.4) may be written as

$$\frac{D^2}{D\tau^2} S^\mu = R^\mu_{\ \nu\lambda\rho} U^\nu U^\lambda S^\rho.$$

Since a slow moving particle $U^\mu = (c, 0, 0, 0) + O(h)$ and the Riemann tensor $R^\mu_{\ \nu\lambda\rho} = O(h)$, this equation has the structure

$$\frac{D^2}{D\tau^2} S^\mu = c^2 \eta^{\mu\sigma} R^{(1)}_{\sigma 00\rho} S^\rho + O(h^2).$$

The Christoffel symbols being of higher order, the covariant derivative may be replaced by ordinary differentiation; this equation at $O(h)$ is

$$\frac{d^2 S^\mu}{dt^2} = \frac{S^\rho}{2} \frac{d^2}{dt^2} h^\mu_{\ \rho}.$$

On the RHS we have used (13.6) and the TT gauge condition of $h_{00} = h_{0\mu} = 0$. The longitudinal component of the separation field S_z is not affected because $h_{3\rho} = 0$ in the TT gauge. For an incoming wave in the “plus” polarization state, the transverse components obey the equations

$$\frac{d^2 S_x}{dt^2} = \frac{S_x}{2} \frac{d^2}{dt^2}(h_+ e^{ikx}), \qquad \frac{d^2 S_y}{dt^2} = -\frac{S_y}{2} \frac{d^2}{dt^2}(h_+ e^{ikx}).$$

These equations, to the lowest order, have solutions

$$S_x(x) = \left(1 + \frac{1}{2} h_+ e^{ikx}\right) S_x(0), \qquad S_y(x) = \left(1 - \frac{1}{2} h_+ e^{ikx}\right) S_y(0)$$

in agreement with the result in (13.37) and (13.38).

(13.3) **$\Gamma^{\mu}_{\nu\lambda}$ and $R^{(2)}_{\mu\nu}$ in the TT gauge**

(a) Christoffel symbols: we give samples of the calculation

$$\Gamma^{1}_{00} = \frac{1}{2}g^{11}(\partial_0 g_{10} + \partial_0 g_{01} - \partial_1 g_{00}) = 0$$

because $h_{10} = h_{01} = h_{00} = 0$ in the TT gauge.

$$\begin{aligned}\Gamma^{1}_{01} &= \frac{1}{2}(1 - \tilde{h}_{11})(\partial_0\tilde{h}_{11} + \partial_1\tilde{h}_{01} - \partial_1\tilde{h}_{10})\\ &= \frac{1}{2}(\partial_0\tilde{h}_{+} - \tilde{h}_{+}\partial_0\tilde{h}_{+}).\end{aligned}$$

(b) Ricci tensor: from what we know of Christoffel symbols having the nonvanishing elements of

$$\begin{aligned}\Gamma^{1}_{10} = \Gamma^{1}_{01} = \Gamma^{0}_{11} &= \frac{1}{2}\partial_0\tilde{h}_{+},\\ \Gamma^{1}_{13} = \Gamma^{1}_{31} = -\Gamma^{3}_{11} &= -\frac{1}{2}\partial_0\tilde{h}_{+}\end{aligned}$$

together with the same terms with the replacement of indices from 1 to 2, we can calculate the second-order Ricci tensor by

$$R^{(2)}_{\mu\nu} = \Gamma^{\alpha}_{\alpha\lambda}\Gamma^{\lambda}_{\mu\nu} - \Gamma^{\alpha}_{\mu\lambda}\Gamma^{\lambda}_{\alpha\nu}.$$

Thus

$$\begin{aligned}R^{(2)}_{00} &= \Gamma^{\alpha}_{\alpha\lambda}\Gamma^{\lambda}_{00} - \Gamma^{\alpha}_{0\lambda}\Gamma^{\lambda}_{\alpha 0}\\ &= 0 - 2\Gamma^{1}_{01}\Gamma^{1}_{10} = \frac{-1}{2}(\partial_0\tilde{h}_{+})^2 = R^{(2)}_{33},\\ R^{(2)}_{11} &= \Gamma^{\alpha}_{\alpha\lambda}\Gamma^{\lambda}_{11} - \Gamma^{\alpha}_{1\lambda}\Gamma^{\lambda}_{\alpha 1}\\ &= 2\Gamma^{1}_{1\lambda}\Gamma^{\lambda}_{11} - \Gamma^{0}_{1\lambda}\Gamma^{\lambda}_{01} - \Gamma^{1}_{1\lambda}\Gamma^{\lambda}_{11} - \Gamma^{3}_{1\lambda}\Gamma^{\lambda}_{31}\\ &= 2\Gamma^{1}_{10}\Gamma^{0}_{11} + 2\Gamma^{1}_{13}\Gamma^{3}_{11} - \Gamma^{0}_{11}\Gamma^{1}_{01} - \Gamma^{1}_{10}\Gamma^{0}_{11}\\ &\quad - \Gamma^{1}_{13}\Gamma^{3}_{11} - \Gamma^{3}_{11}\Gamma^{1}_{31}\\ &= 0 = R^{(2)}_{22}.\end{aligned}$$

(13.4) **Checking the equivalence of (13.62) and (13.63)** We first calculate

$$\begin{aligned}\tilde{I}_{ij}\tilde{I}_{ij} - 2\tilde{I}_{i3}\tilde{I}_{i3} &= \tilde{I}_{i1}\tilde{I}_{i1} + \tilde{I}_{i2}\tilde{I}_{i2} - \tilde{I}_{i3}\tilde{I}_{i3}\\ &= \tilde{I}^{2}_{11} + \tilde{I}^{2}_{22} + 2\tilde{I}_{12}\tilde{I}_{12} - \tilde{I}^{2}_{33}\\ &= 2\tilde{I}_{12}\tilde{I}_{12} - 2\tilde{I}_{11}\tilde{I}_{22},\end{aligned}$$

where we have used

$$\tilde{I}^{2}_{33} = (\tilde{I}_{11} + \tilde{I}_{22})^2 = \tilde{I}^{2}_{11} + \tilde{I}^{2}_{22} + 2\tilde{I}_{11}\tilde{I}_{22}.$$

Thus

$$\begin{aligned}2\tilde{I}_{ij}\tilde{I}_{ij} - 4\tilde{I}_{i3}\tilde{I}_{i3} + \tilde{I}_{33}\tilde{I}_{33} &= 4\tilde{I}_{12}\tilde{I}_{12} - 4\tilde{I}_{11}\tilde{I}_{22} \\ &\quad + \tilde{I}_{11}^2 + \tilde{I}_{22}^2 + 2\tilde{I}_{11}\tilde{I}_{22} \\ &= (\tilde{I}_{11} - \tilde{I}_{22})^2 + 4\tilde{I}_{12}^2\end{aligned}$$

which is the claimed result.

References

Alcock, C. *et al.* (1997). “The MACHO project: Large Magellanic Cloud microlensing results from the first two years and the nature of the galactic dark halo,” *Astrophys. J.*, **486**, 697.

Bennett, C.L. *et al.* (2003). “First year WMAP observations: maps and basic results,” *Astrophys. J.*, suppl. ser., **143**, 1.

Burles, S. *et al.* (2001). “Big-bang nucleosynthesis predictions for precision cosmology,” *Astrophys. J. Lett.*, **552**, L1.

Cheng, T.P. and Li, L.F. (1988). “Resource letter: GI-1 gauge invariance,” *Am. J. Phys.*, **56**, 596.

Cheng, T.P. and Li, L.F. (2000). *Gauge Theory of Elementary Particle Physics: Problems and Solutions*. (Section 8.3), Clarendon Press, Oxford.

Colless, M. (2003). “Cosmological results from the 2dF galaxy redshift survey,” *Measuring and Modeling the Universe*. Carnegie Observatories Astrophysics Series, Vol.2, ed. W.L. Freedman (Cambridge University Press, Cambridge).

Cook, R.J. (2004). “Physical time and physical space in general relativity,” *Am. J. Phys.*, **72**, 214.

Cornish, N.J. *et al.* (2004). “Constraining the topology of the universe,” *Phys. Rev. Lett.*, **92**, 201302.

Cram, T.R. *et al.* (1980). “A complete, high-sensitivity 21-cm hydrogen line survey of M-31,” *Astron. Astrophys.*, Suppl., **40**, 215.

Das, A. (1993). *Field Theory, A Path Integral Approach*. (Section 5.1), World Scientific, Singapore.

de Bernardis, P. *et al.* Boomerang collaboration (2000). “A flat universe from high-resolution maps of the cosmic microwave background radiation,” *Nature*, **404**, 955.

Einstein, A. (1989). *The Collected Papers of Albert Einstein*. Vols 2, 3, and 4, Princeton University Press, Princeton, NJ.

Einstein, A., Lorentz, H.A., Weyl, H., and Minkowski, H. (1952). *The Principle of Relativity—A Collection of Original Papers on the Special and General Theory of Relativity*. Dover, New York.

Ellis, G.F.R. and Williams, R.M. (1988). *Flat and Curved Space-Times*. Clarendon Press, Oxford.

Fixsen, D.J. *et al.* (1996). “The cosmic microwave background spectrum from the full COBE FIRAS data set,” *Astrophys. J.* **473**, 576.

Freedman, W.L. and Turner, M.S. (2003). “Colloquium: measuring and understanding the universe,” *Rev. Mod. Phys.*, **75**, 1433.

Gamow, G. (1970). *My World Line, An Informal Autobiography*. Viking, New York, p 44.

Gott, J.R. *et al.* (2003). “A map of the universe” (arXiv: astro-ph/0310571).

Griest, K. and Kamionkowski, M. (2000). “Supersymmetric dark matter,” *Phys. Rep.*, **333**, 167.

Guth, A.H. (1981). “The inflationary universe: a possible solution to the horizon and flatness problems,” *Phys. Rev. D*, **23**, 347.

Hafele, J.C. and Keating, R.E. (1972). “Around-the-world atomic clocks: observed relativistic time gains,” *Science*, **177**, 168.

Hanany, S. *et al.* (2000) “Constraints on cosmological parameters from MAXIMA-1,” *Astrophys. J. Lett.*, **545**, L5

Kibble, T.W.B. (1985). *Classical Mechanics*. 3rd edn, Longman Press, London.

Krauss, L.M. and Chaboyer, B. (2003). “Age estimates of globular clusters in the Milky Way: constraints on cosmology,” *Science*, **299**, 65.

Landau, L.D. and Lifshitz, E.M. (1975). *The Classical Theory of Fields*. Butterworth-Heinemann/Elsevier, Amsterdam.

Logunov, A.A. (2001). *On the Articles by Henri Poincaré “On the Dynamics of the Electron*,” translated into English by G. Pontecorvo, 3rd edn, JINR, Dubna.

Luminet, J.-P. *et al.* (2003). “Dodecahedral space topology as an explanation for weak wide-angle temperature correlations in the cosmic microwave background,” *Nature*, **425**, 593.

Miller, A.D. *et al.* TOCO collaboration (1999). “A measurement of the angular power spectrum of the CMB from $l = 100$ to 400,” *Astrophys. J. Lett.*, **524**, L1.

Okun, L.B., Selivanov, K.G., and Telegdi, V.L. (2000). “On the interpretation of the redshift in a static gravitational field,” *Am. J. Phys.*, **68**, 115.

Perlmutter, S. *et al.* Supernova Cosmology Project (1999). “Measurements of omega and lambda from 42 high redshift supernovae,” *Astrophys. J.*, **517**, 565.

Peters, P.C. and Mathews, J. (1963). “Gravitational radiation from point masses in a Keplerian orbit,” *Phys. Rev.*, **131**, 435.

Pound, R.V. and Rebka, G.A. (1960). “Apparent weight of photons,” *Phys. Rev. Lett.*, **4**, 337.

Pound, R.V. and Snider, J.L. (1964). “Effects of gravity on nuclear resonance,” *Phys. Rev. Lett.*, **13**, 539.

Riess, A.G. *et al.* High-z Supernova Search Team (1998). “Observational evidence from supernovae for an accelerating universe and a cosmological constant,” *Astron. J.*, **116**, 1009.

Riess, A.G. (2000). “The case for an accelerating universe from supernovae,” *Publ. Astro. Soc. Pac.* **112**, 1284.

Riess, A.G. *et al.* (2001). “The farthest known supernova: support for an accelerating universe and a glimpse of the epoch of deceleration,” *Astrophys. J.,* **560**, 49.

Riess, A.G. *et al.* (2004). “Type Ia Supernova discoveries at $z > 1$ from the Hubble Space Telescope: evidence for past deceleration and constraints on dark energy evolution,” *Astron. J.* (June issue) (arXiv: astro-ph/0402512).

Sadoulet, B. (1999). “Deciphering the nature of dark matter,” *Rev. Mod. Phys.*, **71**, S197.

Schwinger, J. (1986). *Einstein's Legacy—The Unity of Space and Time*. (Chapter 4), Scientific American Books, New York.

Smoot, G.F. *et al.* (1990). "COBE Differential Microwave Radiometers: instrument design and implementation," *Astrophys. J.* **360**, 685.

Smoot, G.F. *et al.* (1992). "Structure in the COBE Differential Microwave Radiometer first year maps," *Astrophys. J.* **396**, L1.

Tolman, R.C. (1934). *Relativity, Thermodynamics and Cosmology*. Clarendon Press, Oxford.

Uhlenbeck, G. (1968). *Introduction to the General Theory of Relativity* (unpublished lecture notes, Rockefeller University).

Weisberg, J.M. and Taylor, J.H. (2003). "The relativistic binary pulsar B1913+16," *Proceedings of Radio Pulsars*, Chania, Crete, 2002 (eds) M. Bailes, *et al.* (ASP. Conf. Series).

White, M. and Cohn, J.D. (2002). "Resource letter: TACMB-1 the Theory of Anisotropies in the Cosmic Microwave Background," *Am. J. Phys.*, **70**, 106.

Wilczek, F. (2004). "Total relativity," *Physics Today*, **57** (No. 4), 10.

Zwiebach, B. (2004). *A First Course in String Theory*. Cambridge University Press.

Picture credits

Fig. 6.5 and book cover: Image from website (http://hubblesite.org/newscenter/newsdesk/archive/releases/2000/07/image/b). Credits: S. Baggett (STScI), A. Fruchter (NASA), R. Hook (ST-ECF), and Z. Levay (STScI).

Fig. 9.6: Image from (de Bernardis *et al*., 2000)

Fig. 13.3: Courtesy of LIGO Hanford Observatory, funded by NSF. Image from website (http://www.ligo-wa.caltech.edu/).

Fig. 13.4: Courtesy of the NAIC — Arecibo Observatory, a facility of the NSF. Image from website (http://www.ligo-wa.caltech.edu/).

Bibliography

This bibliography, by no means an exhaustive listing, contains titles that I have consulted while writing this book. They are arranged so that more recent publications and my personal favorites are placed at the top in each category.

1. **Books at a level comparable to our presentation**

   (a) *General relativity (including cosmology)*

   i. Hartle, J.B., *Gravity: An Introduction to Einstein's General Relativity* (Addison-Wesley, San Francisco, 2003).
   ii. Ohanian, H. and Ruffini, R., *Gravitation and Spacetime*, 2nd edn (Norton, New York, 1994).
   iii. D'Inverno, R., *Introducing Einstein's Relativity* (Oxford U.P., 1992).
   iv. Kenyon, I.R., *General Relativity* (Oxford U.P., 1990).
   v. Schutz, B.F., *A First Course in General Relativity* (Cambridge U.P., 1985).
   vi. Landau, L.D. and Lifshitz, E.M., *The Classical Theory of Fields* (Butterworth-Heinemann/Elsevier, Amsterdam, 1975).

   (b) *Cosmology*

   i. Ryden, B., *Introduction to Cosmology* (Addison-Wesley, San Francisco, 2003).
   ii. Raine, D.J. and Thomas, E.G., *An Introduction to the Science of Cosmology* (Institute of Physics, Bristol, 2001).
   iii. Rowan-Robinson, M., *Cosmology* 4th edn (Oxford U.P., 2003).
   iv. Berry, M.V., *Principles of Cosmology and Gravitation* (Adam Hilger, Bristol, 1989).
   v. Harrison, E., *Cosmology: The Science of the Universe* 2nd edn (Cambridge U.P., 2000).
   vi. Silk, J., *The Big Bang*, 3rd edn (W.H. Freeman, New York, 2000).
   vii. Bergstrom, L. and Goobar, A., *Cosmology and Particle Astrophysics* (Wiley, New York, 1999).

2. **Books at a more advanced level**

   (a) *General relativity (including cosmology)*

   i. Misner, C., Thorne, K., and Wheeler, J.A., *Gravitation* (W.H. Freeman, New York, 1970).
   ii. Weinberg, S., *Gravitation and Cosmology* (Wiley, New York, 1972).
   iii. Stephani, H. *General Relativity* 2nd edn (Cambridge U.P., 1990).
   iv. Wald, R.M., *General Relativity* (Chicago U.P., 1984).

   (b) *Cosmology*

   i. Peacock, J.A., *Cosmological Physics* (Cambridge U.P., 1999).

ii. Peebles, P.J.E., *Principles of Physical Cosmology* (Princeton U.P., 1993).

iii. Kolb, E.W. and Turner, M.S., *The Early Universe* (Addison-Wesley, San Francisco, 1990).

3. General interest and biographical books

i. Thorne, K.S., *Black Holes & Time Warps: Einstein's Outrageous Legacy* (Norton, New York, 1994).

ii. Will, C., *Was Einstein Right?—Putting General Relativity to the Test* (Basic Books, New York, 1986).

iii. Pais, A., *Subtle is the Lord... The Science and Life of Albert Einstein* (Oxford U.P., 1982).

iv. Weinberg, S., *The First Three Minutes* (Basic Books, New York, 1972).

v. Schwinger, J., *Einstein's Legacy—The Unity of Space and Time* (Scientific American Books, New York, 1986).

vi. Guth, A.H., *The Inflationary Universe* (Addison-Wesley, San Francisco, 1997).

vii. Green, B., *The Elegant Universe* (Norton, New York, 1999).

viii. Smolin, L., *Three Roads to Quantum Gravity* (Basic Books, New York, 2001).

ix. Goldsmith, D., *The Runaway Universe* (Perseus, Cambridge MA, 2000).

x. Zee, A., *Einstein's Universe: Gravity at Work and Play* (Oxford U.P., 2001).

xi. French, A. (ed.), *Einstein—A Centenary Volume* (Harvard U.P., 1979).

xii. Howard, D. and J. Stachel (eds), *Einstein and the History of General Relativity* (Birkhäuser Boston, 1989).

Index