

Graduate Texts in Physics

Ronald J. Adler

General Relativity and Cosmology

A First Encounter



Springer

Graduate Texts in Physics

Series Editors

Kurt H. Becker, NYU Polytechnic School of Engineering, Brooklyn, NY, USA

Jean-Marc Di Meglio, Matière et Systèmes Complexes, Bâtiment Condorcet,
Université Paris Diderot, Paris, France

Morten Hjorth-Jensen, Department of Physics, Blindern, University of Oslo, Oslo,
Norway

Bill Munro, NTT Basic Research Laboratories, Atsugi, Japan

William T. Rhodes, Department of Computer and Electrical Engineering and
Computer Science, Florida Atlantic University, Boca Raton, FL, USA

Susan Scott, Australian National University, Acton, Australia

H. Eugene Stanley, Center for Polymer Studies, Physics Department, Boston
University, Boston, MA, USA

Martin Stutzmann, Walter Schottky Institute, Technical University of Munich,
Garching, Germany

Andreas Wipf, Institute of Theoretical Physics, Friedrich-Schiller-University Jena,
Jena, Germany

Graduate Texts in Physics publishes core learning/teaching material for graduate- and advanced-level undergraduate courses on topics of current and emerging fields within physics, both pure and applied. These textbooks serve students at the MS- or PhD-level and their instructors as comprehensive sources of principles, definitions, derivations, experiments and applications (as relevant) for their mastery and teaching, respectively. International in scope and relevance, the textbooks correspond to course syllabi sufficiently to serve as required reading. Their didactic style, comprehensiveness and coverage of fundamental material also make them suitable as introductions or references for scientists entering, or requiring timely knowledge of, a research field.

More information about this series at <http://www.springer.com/series/8431>

Ronald J. Adler

General Relativity and Cosmology

A First Encounter

Ronald J. Adler
Department of Physics and Astronomy
San Francisco State University
San Francisco, CA, USA

Gravity Probe B Mission
Hansen Experimental Physics Laboratory
Stanford University
Stanford, CA, USA

ISSN 1868-4513

ISSN 1868-4521 (electronic)

Graduate Texts in Physics

ISBN 978-3-030-61573-4

ISBN 978-3-030-61574-1 (eBook)

<https://doi.org/10.1007/978-3-030-61574-1>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2021

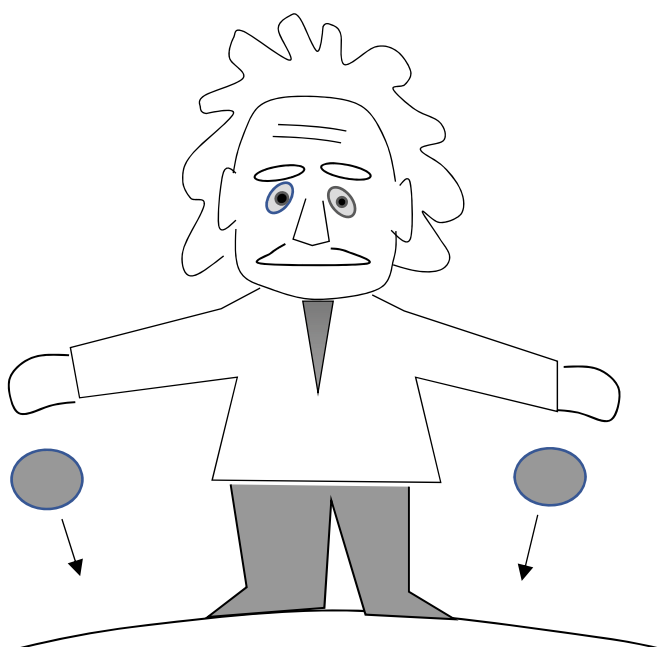
This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Cover image: © Paulista/stock.adobe.com

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland



Preface

Four truly profound questions have always permeated science and served as its basis: what is matter, what is the universe, what is life, and what is thought. The twentieth century has been extraordinary in that three of these have been at least partially answered. The answers in brief and broad outline are: matter is made of quarks and leptons and the quantum gauge fields that hold them together; the universe is an isotropic and homogeneous expanding curved spacetime that is dominated on the cosmological scale by gravity; life is a mechanism whereby the molecular polymer deoxyribonucleic acid (DNA) makes more DNA from its environment.

We have not been so successful concerning the nature of thought; indeed some people are of the opinion that almost no real progress has been made. One rather entertaining view is that thought is an illusion and we do not really think at all. That is to say we merely think that we think. Perhaps it is good that there remains such a field with so much to be explored in our future mental adventures.

While the nature of matter and life are part of the undergraduate curriculum in physics and biology, the nature of the universe is often left for a graduate course on general relativity, and cosmology is often treated only briefly at the end of the course. Some universities offer an undergraduate course on cosmology not based on general relativity, but of course this is no substitute for a more complete treatment. There is no good reason that general relativity and cosmology should not be studied by undergraduates as well as beginning graduate students. This book is thus directed primarily at beginning graduate students but also at advanced and confident undergraduate students. The mathematics needed is only a short step beyond vector and matrix analysis, the physical concepts are simpler than those of quantum mechanics, and the subjects have become very mainstream. Such topics as black holes, dark matter, the shape of the universe, the big bang, the primordial fireball, and the ultimate fate of the universe can be appreciated and understood by almost anyone with an undergraduate physics background. Some of the appendices should help undergraduates and others with gaps in their background.

General relativity theory began in the early twentieth century in the borderland between physics and mathematics. After the initial confrontation of the theory with the three classic tests (red shift, Mercury perihelion shift, deflection of starlight), there was little contact between theory and observation until the last half of the century. But then the discovery of the cosmic microwave background radiation made it clear that the theory had much to offer for describing the evolving universe. Since then the field of observational cosmology has blossomed, using many different approaches to measuring the properties of the universe on a large scale. Theoretical cosmology has naturally blossomed with it and the combination of observation and theory has resulted in the present standard model of cosmology, the lambda cold dark matter or Λ CDM model. It is fair to say that there is now no more active area in fundamental physics than cosmology.

But we must not underestimate the progress in relativity theory and observation for other basic systems, notably neutron stars and black holes. The agreement between black hole theory based on the Kerr metric and diverse observations is one of the most impressive successes in physics. This is most relevant now that it has become apparent how important supermassive black holes are for the structure and evolution of the universe.

Another truly extraordinary prediction of general relativity has been verified with the observation of gravitational waves. The first waves detected were generated by binary black hole and neutron star mergers, using the LIGO and Virgo detectors. The detection required a century of thought and decades of experimental effort. Certainly, the connection of the two extraordinary predictions of relativity theory, black holes and gravitational waves, is most impressive and gratifying. The future promises to be even more interesting since gravitational waves are an entirely new observational window on the cosmos, and there is no way to predict what they might reveal.

Clearly, the frontier of fundamental physics research has now shifted to the large end of the distance scale, the universe. But our understanding of the universe requires also an understanding of the small end of the distance scale, most notably in our study of the early universe. The thriving field now called particle astrophysics and cosmology (PAC) did not even exist until almost the twenty-first century but is now the center of much frontier research.

A remarkable fact concerning the detection of neutron star mergers and the gravitational waves they emit is worth noting here; the kilonovas that are the end result of the mergers are the source of much of the heavier elements we observe in the universe, including the matter that makes up our planet and notably—ourselves.

The purpose of this book is to introduce the reader to general relativity theory and all that it can tell us about the universe. It is intended to be as clear, simple, and brief as possible, and as rigorous as reasonable. It is divided into four somewhat independent parts that might be considered separate volumes.

Part I is a brief review of special relativity; most physics students will have studied special relativity in other courses and may skim easily over this part, but it can serve as a brief introduction for others.

Part II provides mathematical background regarding Riemann space and the vectors and tensors that inhabit it. It uses the ideas and notation of the component or classic approach to tensors but also includes discussion of the more modern ideas and notation of the intrinsic abstract view of tensors.

Part III gives a view of basic general relativity as a theory of gravity and the geometric ideas that underlie it; it includes chapters on gravitational waves and black holes and includes a brief heuristic discussion on Hawking radiation.

Part IV is a survey of relativity theory as used in cosmology; this book is mostly about theory and its mathematical basis, but Part IV includes, of necessity, a fair amount of material on the experimental and observational work being done or being planned. Also of necessity, the material on the observational work is far from complete, but is intended to provide a start and references for anyone interested in pursuing it further and perhaps becoming a researcher.

My prime target readers are early graduate students or advanced and confident undergraduate physics students. A student with the usual mathematical background in calculus and vectors and matrices and a physics background in mechanics and electromagnetism should be able to handle the material without much outside reading. Nominally, the entire book should be readable in a two semester or a two or three quarter course.

Thanks are due to many people who helped in the writing of this book. Much of its content is based on my teaching as an adjunct professor at San Francisco State University and work done on the Gravity Probe B mission at Stanford University. My SFSU relativity classes have given me much feedback and corrected numerous errors. At Stanford and at SLAC National Accelerator Laboratory my colleagues Robert Wagoner, James Bjorken, Francis Everitt, Alex Silbergleit, Pisin Chen, David Santiago, and John Berberian have provided many interesting ideas and discussions. Fred Martin patiently proof-read, criticized, and improved the early notes. James Overduin encouraged me to revise and expand the notes for publication and provided thought-provoking comments.

San Francisco/Stanford, USA

Ronald J. Adler
email: gyreron@gmail.com

Contents

Part I Special Relativity in Review

1	A Brief Stroll in Special Relativity	3
1.1	The Trouble with Absolute Time	3
1.2	The Simplest Lorentz Transformation	5
1.3	Some Elementary Properties and Applications	8
2	Lorentz Transformations	13
2.1	The Lorentz Group	13
2.2	Four-Vectors and Tensors	15
3	The Motion of Particles	19
3.1	Energy and Momentum	19
3.2	Acceleration	23
3.3	Accelerated Motion	25
3.4	Curves and Arc Lengths	27

Part II Vectors and Tensors

4	Riemann Spaces and Tensors	33
4.1	Riemann Spaces	33
4.2	Vectors, Component View	38
4.3	Vectors and 1-Forms, Abstract View	40
4.4	Tensors, Component View	44
4.5	Tensors, Abstract View	48
4.6	Tetrads and n -Trads	49
4.7	Volume Elements	52
5	Affine Connections and Geodesics	59
5.1	Affine Connections, Component View	59
5.2	Transformation of the Affine Connections	61
5.3	Parallel Displacement	64

5.4	Geodesics as Self-parallel Curves	67
5.5	Geodesics as Extremum Curves	69
5.6	Affine Connections, Abstract View	73
6	Tensor Analysis	81
6.1	Covariant Derivatives, Component View	81
6.2	Covariant Derivatives, Abstract View	85
6.3	The Divergence and Laplacian	87
 Part III General Relativity		
7	Classical Gravity and Geometry	95
7.1	Newtonian Gravity	95
7.2	The Equivalence Principle	98
7.3	Gravity as a Geometric Phenomenon	102
8	Curved Space and Gravity	109
8.1	Curved Space and the Riemann Tensor	109
8.2	Symmetries of the Riemann Tensor	113
8.3	The Einstein Equations for the Gravitational Field in Vacuum	115
8.4	The Non-vacuum Field Equations	117
8.5	The Intrinsic Signature of Gravity	121
9	Spherically Symmetric Gravitational Fields	125
9.1	The Schwarzschild Solution	125
9.2	Orbit of a Planet	129
9.3	Deflection of Light	134
9.4	Observational Tests of General Relativity	137
10	Black Holes and Gravitational Collapse	141
10.1	Schwarzschild Black Hole	141
10.2	Null Surfaces	145
10.3	Stellar Evolution, Very Briefly	148
10.4	Collapse of a Dust Star	149
10.5	Spinning Black Holes and the Kerr Metric	150
10.6	Black Holes in the Real Universe	152
10.7	Hawking Radiation from a Black Hole	153
11	Linearized General Relativity and Gravitational Waves	159
11.1	The Field Equations of the Linearized Theory	159
11.2	The Classical Limit	163
11.3	Gravitational Plane Waves	164
11.4	Motion of Test Bodies in Gravitational Waves	168
11.5	Gravitational Wave Sources	171
11.6	Detection of Gravitational Waves	179

Part IV Cosmology

12	The Einstein Field Equations for Cosmology	193
12.1	The Field Equations and Energy-Momentum Conservation	193
12.2	Field Equations and the Cosmic Fluid Source	195
12.3	The Cosmological Constant as Vacuum or Dark Energy	198
12.4	Summary	200
13	Cosmological Preliminaries	203
13.1	Basic Observations and Assumptions	203
13.2	The Cosmological FLRW Metric	207
13.3	Consequences of the Metric	211
13.4	De Sitter Space	217
14	The Dynamical Equations of Cosmology	223
14.1	The Einstein Field Equations for Cosmology	223
14.2	Critical Density and the Shape of the Universe	225
14.3	Observed Dark Matter and Dark Energy Densities	226
14.4	Evolution of Cosmic Fluid Constituents	227
14.5	The Friedmann Master Equation	230
15	Solutions for the Present Universe	233
15.1	The Positive Cosmological Constant	233
15.2	Complete Solution of the Friedmann Master Equation	234
15.3	Cosmological Constant Dominance	234
15.4	Matter Dominance	236
15.5	The Λ CDM Universe	238
16	Some Properties of the ΛCDM Universe	247
16.1	Diverse Cosmological Observations	247
16.2	Cosmological Parameter Values	251
16.3	The Hubble Function and the Age of the Universe	252
16.4	Transition Time for Matter to Dark Energy Dominance	253
16.5	Density Ratios and the Shape of the Universe	254
16.6	Horizons and the Size of the Observable Universe	257
16.7	Conformal Time	259
17	Earlier Times and Radiation	263
17.1	Radiation and Temperature in Earlier Times	263
17.2	The Scale Factor and Basic Properties of the Radiation Era	267
17.3	The Isotropic CMB and the Horizon Puzzle	269
17.4	The Anisotropies of the CMB	270
18	A Brief Historical Overview of the Universe	275
18.1	Overview	275
18.2	Condensation of Stars and Galaxies	277
18.3	Condensation of Atoms	277

18.4	Condensation of Nuclei	278
18.5	Condensation of Nucleons	278
18.6	Inflation	279
18.7	Planck Era	279
19	Inflation and Some Questions	281
19.1	Basic Ideas of Inflation	281
19.2	Inflation Via Scalar Fields	284
19.3	Origin of Structure	286
19.4	The Physical Nature of Dark Energy	290
19.5	The Physical Nature of Dark Matter	291
19.6	The Planck Era and Quantum Physics	292
	References	303
	Index	309

Part I

Special Relativity in Review

Most undergraduate physics students have studied special relativity by the time they are seniors and are familiar with its basic ideas, such as the Lorentz transformation and length contraction and time dilation. However, they may not have been fully exposed to the geometric point of view of spacetime and may not appreciate the formalism and power of four vectors and tensors. Part I has been included for such students, as well as for readers without a background in special relativity. Those confident with their understanding may of course skip or skim this part.

Chapter 1 is a simple review of what most students encounter in a modern physics course, a discussion of time in a universe with a constant velocity of light, and the consequences of the relativity of time such as time dilation and length contraction.

Chapter 2 uses more sophisticated mathematics in a discussion of the Lorentz group and vectors and tensors in spacetime. One important goal is to prepare the reader for the more general vector and tensor algebra and analysis used later in Part II.

Chapter 3 is devoted to the motion of particles, their energy and momentum and acceleration, and emphasizes the geometric view of motion in spacetime. In particular It demonstrates that special relativity is not limited to motion at constant velocity, a misconception that is sometimes encountered.

Chapter 1

A Brief Stroll in Special Relativity



Abstract This chapter is a short review of what students generally encounter in a modern physics course: a discussion of time in a universe with a constant velocity of light, and the important consequences of the relativity of time such as length contraction and time dilation.

1.1 The Trouble with Absolute Time

The story of the discovery of special relativity is one of the most interesting in physics, and is covered in many books, including several by Einstein (Einstein 1923, 1934; Bergmann 1942; Rindler 1969; Weaver 1987). Accordingly we will here discuss only very briefly the ideas which led Einstein to special relativity.

In the late nineteenth century the two great theories of physics were Newton's mechanics and gravitational theory, and Maxwell's electromagnetism. It was widely believed that there might be no more basic physical theories to be discovered: quantum mechanics was of course decades in the future. However there was a flaw in the combination of these two theories, inherent in the classical concept of time. Mechanics was based on absolute time; as Newton phrased it in the *Principia*, "Absolute, true, and mathematical time, of itself, and from its own nature, flows equally without reference to anything external, and by another name is called duration: relative, apparent, and common time, is some sensible and external (whether accurate or unequable) measure of duration by the means of motion, which is commonly used instead of true time; such as an hour, a day, a month, a year."

The transformation between Cartesian reference frames in uniform motion, called the Galilean transformation, is based on the notion of absolute time, and was universally accepted in the nineteenth century. For motion along the x direction the situation is shown in Fig. 1.1; the primed system moves past the unprimed system at velocity v , with the origins coinciding at time zero.

The Galilean transformation between the two systems is

$$x' = x - vt, \quad y' = y, \quad z' = z, \quad t' = t = \text{absolute time.} \quad (1.1)$$

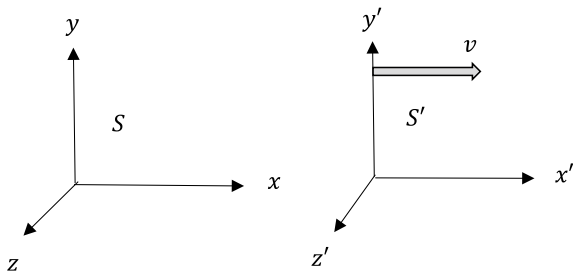


Fig. 1.1 Reference frames and coordinate systems in relative motion in the x direction

If a body moves with velocity u in the x direction in system S then it will have a velocity in system S' given by differentiating this with respect to the absolute time,

$$u' = \frac{dx'}{dt} = \frac{dx}{dt} - v = u - v, \quad u = u' + v. \quad (1.2)$$

That is the velocities u' and v simply add to give $u' + v$. You may easily convince yourself that the general vector expression for the addition of velocities must be

$$\vec{u} = \vec{u}' + \vec{v}. \quad (1.3)$$

The invariance of Newton's second law under the transformation is evident since the relative velocity is a constant and the acceleration is then the same in both systems. This is the basis of Galilean invariance: Newton's laws and the behavior of mechanical systems are the same in all uniformly moving reference frames.

At the end of the nineteenth century Maxwell's electromagnetism was generally accepted, partly because it predicted the correct velocity for light, $c = 1/\sqrt{\mu_0\epsilon_0} = 2.9979 \times 10^8$ m/s, and it even predicted the existence of radio waves. But this implied an interesting fact, that the velocity of light, according to (1.3), should be different in different frames. Thus Maxwell's equations should somehow be different in different frames, either in the value of $\mu_0\epsilon_0$ or in their mathematical structure. The conventional viewpoint was that the equations were valid and c had the indicated value in one special frame, that in which the supposed medium that supported light waves, the luminiferous ether, was *at rest*.

This viewpoint was apparently self-consistent. The problem came when experimenters searched for evidence of the ether and of the velocity of the earth through the ether and did not find it. The best known such experiment was that of Michelson and Morley, which we will not discuss here since it is discussed in many books (Taylor 1963). A number of phenomenological explanations were proposed to explain the failure to observe effects of the ether but were largely forgotten when Einstein presented his explanation in terms of the theory of special relativity.

Einstein's approach was to assume that the Maxwell equations were valid and that the speed of light was the same in all inertial systems, and then to rethink the whole question of space and time, based on the constancy of the speed of light. The result

was that he abandoned Newton's absolute time and developed the special theory of relativity.

1.2 The Simplest Lorentz Transformation

Einstein's 1905 approach to special relativity was based on the following two postulates:

- I. *The analytical form of physical laws is the same in all inertial reference frames as described by systems of Cartesian coordinates.*
- II. *The speed of light in vacuum is a universal constant.*

Postulate (I) is a criterion of elegance, while (II) was supported by experiments done before 1905, such as that of Michelson and Morley, and is now verified to very high accuracy.

We want to derive now a transformation of the space coordinates *plus time*, to replace the Galilean transformation discussed above, but in which the velocity of light is the same in both systems. This is called a Lorentz transformation; due to its fundamental importance our derivation will be detailed and based on the most elementary assumptions (Sard 1970).

To begin we modify the Galilean transformation (1.1) in as simple a way as we can. First, we suppose that y and z are not changed, that is $y' = y$ and $z' = z$ (You should think about this a little). We next assume that time may be different in the two systems, and that the transformation is linear in x and t . That is we assume

$$ct' = a_{11}ct + a_{12}x, \quad x' = a_{21}ct + a_{22}x. \quad (1.4a)$$

In equivalent matrix form,

$$\begin{pmatrix} ct' \\ x' \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} ct \\ x \end{pmatrix}, \quad A(v) \equiv \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}. \quad (1.4b)$$

The matrix elements a_{ij} must, of course, depend only on the velocity v . The notable property of this transformation is that time is allowed to be different in the two systems, which is the fundamental break with classical ideas made by Einstein. It is this which allows c to be a universal constant. The use of ct instead of t in (1.4a) is for dimensional convenience, since ct and x both have dimensions of distance. There are 4 parameters in the transformation matrix A , which we must determine. We will make four physical demands based on the above two postulates that determine them uniquely.

Demand I. We can describe the origin of the system S' in terms of both coordinate systems. In the primed coordinates it is given by $x' = 0$ and in the unprimed coordinates it is given by $x = vt$. This is simply the statement that S' moves at velocity v relative to S . We use (1.4a) to express $x' = 0$ as

$$x' = a_{21}ct + a_{22}x = 0. \quad (1.5)$$

Then we substitute $x = vt$ to obtain

$$a_{21}ct + a_{22}vt = 0, \quad (1.6)$$

and thus

$$a_{21} = -(v/c)a_{22}, \text{ from Demand 1.} \quad (1.7)$$

Demand 2. We can repeat the above argument from the opposite perspective, that is by noting that S moves at $-v$ with respect to S' . The origin of S corresponds to $x = 0$ in terms of unprimed coordinates, and to $x' = -vt'$ in terms of primed coordinates. Then $x = 0$ substituted in (1.4a) gives

$$ct' = a_{11}ct, \quad x' = a_{21}ct. \quad (1.8)$$

Substitution of (1.8) into $x' = -vt'$ tells us that

$$a_{21}ct = -va_{11}t, \quad (1.9)$$

so we find from (1.9) and (1.7)

$$a_{21} = -(v/c)a_{11} \quad \text{and} \quad a_{22} = a_{11}, \text{ from Demand 2.} \quad (1.10)$$

Demand 3. The third demand is much deeper; it has to do with the velocity of light in the two systems. Suppose a very brief pulse of light is emitted as the origins of the two systems coincide, at $x = x' = 0$. Then by the postulate II, that the speed of light be the same in the two systems, the pulse will be at $x = ct$ in S and at $x' = ct'$ in S' . We write the second, $x' = ct'$, using the transformation (1.4a) as

$$a_{21}ct + a_{22}x = a_{11}ct + a_{12}x. \quad (1.11)$$

Then we use the first, $x = ct$, to infer that

$$a_{21}ct + a_{22}ct = a_{11}ct + a_{12}ct, \quad \text{so} \quad a_{21} = a_{12}. \quad (1.12)$$

Combining this with (1.7) and (1.10) we have

$$a_{12} = a_{21} = -(v/c)a_{11}, \text{ from Demand 3.} \quad (1.13)$$

Before we make the fourth demand let us collect our results. From the above three demands we see that all the elements of the transformation matrix are determined except a_{11} and the transformation matrix may be written as

$$A(v) = a_{11} \begin{pmatrix} 1 & -v/c \\ -v/c & 1 \end{pmatrix}. \quad (1.14)$$

Demand 4. Only the parameter a_{11} remains to be determined. The transformation matrix $A(v)$ transforms from S to S' . Thus the inverse transformation matrix $A(v)^{-1}$ transforms from S' to S . But we could clearly reverse the roles of the two systems and see that the transformation matrix $A(-v)$ should also transform from S' to S . Therefore we have two expressions for the inverse transformation and see that $A(v)^{-1}$ must be the same as $A(-v)$. These matrices, with the dependence on v stated explicitly, are easily gotten from (1.14)

$$A(v)^{-1} = \left(\frac{1}{a_{11}(v)} \right) \begin{pmatrix} 1 & v/c \\ v/c & 1 \end{pmatrix} \frac{1}{1 - v^2/c^2}, \quad (1.15)$$

$$A(-v) = a_{11}(-v) \begin{pmatrix} 1 & v/c \\ v/c & 1 \end{pmatrix}.$$

Since these must be equal we get a simple relation for $a_{11}(v)$

$$a_{11}(-v)a_{11}(v) = \frac{1}{1 - v^2/c^2}. \quad (1.16)$$

As part of Demand 4 we also ask that a_{11} depend only on the magnitude of the velocity rather than its direction, so that $a_{11}(-v) = a_{11}(v)$, and thus obtain

$$a_{11} = 1/\sqrt{1 - v^2/c^2} \equiv \gamma, \quad \text{from Demand 4.} \quad (1.17)$$

We will justify the demand that a_{11} depend only on v^2 further below when we discuss the rate of a moving clock; the rate of such a clock must be independent of its direction of motion to be consistent with the isotropy of space. See the time dilation expression (1.20) and Exercise 1.6.

Let us summarize the important result of this section. The fundamental Lorentz transformation for one space dimension, written in terms of parameters β and γ , is

$$ct' = \gamma ct - \beta \gamma x, \quad x' = \gamma x - \beta \gamma ct, \quad (1.18)$$

$$A(v) = \gamma \begin{pmatrix} 1 & -\beta \\ -\beta & 1 \end{pmatrix}, \quad \text{Lorentz transformation matrix,}$$

where the ubiquitous parameters β and γ are defined as

$$\beta \equiv v/c, \quad \gamma \equiv 1/\sqrt{1 - \beta^2}. \quad (1.19)$$

This is the famous Lorentz transformation for motion in the x direction; γ is termed the Lorentz contraction factor, which we will often call simply the γ factor. There is a wealth of interesting physics in this transformation, a little of which we will discuss next.

1.3 Some Elementary Properties and Applications

Many of the most interesting results of special relativity theory can be obtained using only the simple Lorentz transformation above (Taylor 1963). We will give a rather cursory discussion of some of the more important features, appropriate to a review: time dilation of a moving clock, length contraction of a moving rod, and the Doppler shift of light emitted by a moving object. The interested reader may consult the references for much more material.

First note that the Lorentz transformation contains the factor γ , which is greater than 1. If γ is not to be infinite or imaginary then the velocity parameter β must be less than 1; thus systems and objects cannot move faster than c , a famous result of relativity.

Example 1.1 What is “fast”? It is clear from the above, and we will soon see further, that the γ factor is a good indicator of when relativistic effects become important. For zero velocity it is equal to 1, and for velocity equal to c it is infinite. We may, somewhat arbitrarily, take the velocity at which $\gamma = 1.1$ to be fast, that is for which relativistic effects are of order of 10%. Then “fast” means $\beta = 1/\sqrt{1 - 1/\gamma^2} = 0.42$ or $v = 1.25 \times 10^7$ m/s. It turns out that this implies that classical mechanics is rather accurate for surprisingly large velocities.

Time dilation in a moving system is an effect peculiar to relativity, which distinguishes it sharply from classical theory with its absolute time. Suppose a clock at rest at the origin in the moving system S' ticks at $t' = 0$ and again at $t' = \Delta t'$. Then in the system S , where we suppose our lab to be, it is seen to tick at $t = 0$ at $x = 0$ and again at $t = \Delta t$ at $x = v\Delta t$. With the Lorentz transformation in (1.18) we may relate these time intervals,

$$c\Delta t' = \gamma c\Delta t - \beta\gamma\Delta x = \gamma c\Delta t - \beta\gamma v\Delta t = c\Delta t/\gamma \quad \text{or} \quad \Delta t = \gamma\Delta t'. \quad (1.20)$$

Thus, since $\gamma \geq 1$, the moving clock appears to run slower as seen in the lab in S . We refer to the system in which a clock is at rest as its *rest system* or *proper system* or *rest frame*. Time in the proper system is usually called *proper time* and often denoted by τ .

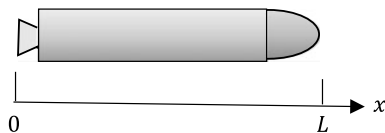


Fig. 1.2 The rocket nose is at $x = L$ and the tail at $x = 0$ at $t = 0$ in our lab frame

Example 1.2 Muons have a lifetime of about $2 \mu\text{s}$ in their rest frame. In a universe with absolute time they could travel only about 600 m before decaying if moving at nearly c . In fact they have been observed to travel many km. The “little clock inside the muon” must indeed run slow.

Length contraction is one of the best-known properties of relativity. It involves two facets of the theory—the definition of length and the relativity of simultaneity. Suppose an object such as a rocket ship is at rest in system S' with its tail at $x' = 0$ and its nose at $x' = L_p$, which of course we call its *proper length*. In our lab in S we observe the ship pass by so that at $t = 0$ its tail is at $x = 0$ and its nose is at $x = L$, which we call its length in the lab system. This is shown in Fig. 1.2. From this the Lorentz transformation (1.18) gives a relation between L and L_p ,

$$L_p = x' = \gamma x - \beta \gamma ct = \gamma x = \gamma L, \quad L = L_p / \gamma, \quad (1.21)$$

since $t = 0$ in the lab. That is, in the lab we observe the moving rocket to be shorter than its length in the rest or proper frame. Note that this nonintuitive result is obtained since the positions of nose and tail are observed simultaneously at $t = 0$ in the lab, a fundamental part of the definition of length implied in the above. Observers in the rocket’s proper frame will not consider the measurement in the lab frame to be valid since they will see a time difference between the nose and tail measurements of

$$ct' = \gamma ct - \beta \gamma x = -\beta \gamma L \neq 0. \quad (1.22)$$

That is the simultaneous measurements of nose and tail positions in the lab are not simultaneous in the proper system; *simultaneity* is relative to the system. This was one of Einstein’s great insights which led to special relativity.

Example 1.3 Do objects *visually* appear to be contracted according to (1.21)? They do not. The definition of length in the above example does not involve visual appearance. Consider a rocket moving directly toward us on the x axis. We ask where we *see* the nose and tail of the rocket as it moves toward us, and take this as the definition of its *visual length*. Figure 1.3 shows the rocket at two times; one photon from the tail (T) and one from the nose (N) of the

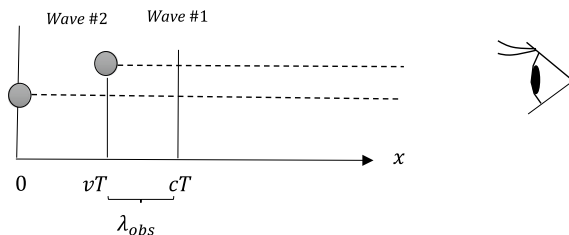


Fig. 1.4 The light source at two different times. It moves directly toward the observer

We can relate this to the period T_p in the proper frame with the time dilation equation (1.20) to find

$$\lambda_{\text{ob}} = (1 - \beta)cT = (1 - \beta)\gamma cT_p = (1 - \beta)\gamma\lambda_p, \quad (1.27)$$

where $\lambda_p = cT_p$ is the wavelength in the proper frame; the contribution of relativity to this expression is the factor of γ . If the source moves toward us we therefore observe a shorter wavelength, that is a blue shift. You should think through the argument for a source that moves away from the observer and verify that the sign of the velocity in (1.27) changes and one observes a red shift. The general case in which the source moves at an arbitrary angle is not difficult; see Exercise 1.4.

There is another velocity measure, called *rapidity*, that is often more useful than β . Notice that under successive Lorentz transformations the velocities are not additive; that is $A(\beta_1)A(\beta_2) \neq A(\beta_1 + \beta_2)$ (see Exercise 1.3). Rapidity is defined so that the rapidities of successive Lorentz transformations do add; specifically, we define rapidity θ by $\beta = \tanh \theta$, so the Lorentz transformation (1.18) may be written as

$$A(\theta) = \begin{pmatrix} \cosh \theta & -\sinh \theta \\ -\sinh \theta & \cosh \theta \end{pmatrix}, \quad \beta = \tanh \theta, \quad \gamma = \cosh \theta, \quad (1.28)$$

(see Exercise 1.5). It is easy to verify that $A(\theta_1)A(\theta_2) = A(\theta_1 + \theta_2)$. That is, rapidity is an additive measure, as desired (see Exercise 1.3 for further motivation for the definition). We will find this property useful when we discuss accelerated motion in Chap. 3.

Exercises

- 1.1 At the SLAC National Accelerator Center electrons were accelerated to have $\gamma = 4 \times 10^4$ so they were moving at nearly c . To see how near set $\beta = 1 - \varepsilon$ then calculate ε approximately.
- 1.2 Suppose that you will live for another 100 years or so. The universe is about 10 billion light years across—as we will later discuss. About how fast must you move to cross it in your lifetime?
- 1.3 Velocities behave rather differently in relativity than in classical mechanics. Study the addition of velocities by considering 3 inertial systems as follows. We are at rest in S , system S' moves with respect to us at β_1 , while system

S'' moves with respect to system S' at β_2 . To see how fast system S'' moves with respect to us multiply the individual Lorentz transformations using the matrix representation (1.18). The product will be a Lorentz transformation, that is $A(\beta_1)A(\beta_2) = A(\beta)$ with $\beta = (\beta_1 + \beta_2)/(1 + \beta_1\beta_2)$. This is the addition law for velocities. Notice that for small velocities it agrees with the classical law, while for both velocities approaching c the total velocity remains less than c , approaching c from below.

- 1.4 Derive the general expression for the Doppler shift,

$$\lambda_{\text{ob}} = (1 - \beta \cos \theta) \gamma \lambda_p,$$

where θ is the angle between the velocity of the source and a line between the source and the observer. Notice that even for $\theta = 90^\circ$ there is a shift; this is called the transverse Doppler shift and is not present in the classical theory (Taylor 1963).

- 1.5 We defined rapidity as a convenient alternative measure of velocity in the text. It can also be motivated if we consider rotations in 2 dimensions as an analog. The usual matrix representation of a rotation is

$$R(\theta) = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}.$$

Show that angles are additive, that is $R(\theta_1)R(\theta_2) = R(\theta_1 + \theta_2)$.

However we could also measure the rotation by the tangent of θ , call it α . In this case the rotation matrix would be

$$R(\alpha) = \frac{1}{\sqrt{1 + \alpha^2}} \begin{pmatrix} 1 & -\alpha \\ \alpha & 1 \end{pmatrix}.$$

Show that with the α measure the rotations are not additive, that is $R(\alpha_1)R(\alpha_2) \neq R(\alpha_1 + \alpha_2)$. We may conclude that α is not a very convenient rotation measure. Notice the similarity between the matrix above and the Lorentz transformation matrix in (1.18); that is what leads to the hyperbolic definition of rapidity.

- 1.6 One could have a solution to (1.16) with $a_{11}(v) = 1/(1 - v/c)$. Investigate how this would affect the rate of a moving clock according to (1.20) and show that it is not compatible with the isotropy of space.

Chapter 2

Lorentz Transformations



Abstract This chapter uses the mathematics of matrices to discuss the Lorentz transformation and vectors and tensors in spacetime. One important goal is to prepare the reader for the more general vector and tensor algebra and analysis to be used in Part II.

2.1 The Lorentz Group

We have obtained the Lorentz transformation for motion in the x direction and discussed some elementary applications. Now we are going to look at such transformations from a more sophisticated mathematical viewpoint, and with a more elegant notation (Schutz 2009). This chapter is intended to orient you towards the geometric viewpoint of general relativity, and to show that the notation can do much of the algebraic work for you. Only cartesian coordinates will be used in this chapter.

We will first derive a more general definition of a Lorentz transformation. Recall that special relativity is based on the following principles (Schwartz 1968).

- I. *The analytical form of physical laws is the same in all inertial reference frames as described by systems of Cartesian coordinates.*
- II. *The speed of light in vacuum is a universal constant.*

A more sophisticated way to state principle II is that we wish to make the equation of an expanding spherical wave front of light *invariant* under the relevant transformation of the space and time coordinates. We write the wave front as

$$c^2 t^2 - \vec{x}^2 = 0, \quad (2.1)$$

and show a picture in Fig. 2.1 with the z coordinate suppressed. Because of the shape of the surface in this picture it is called a light cone. Events in an inertial system are points in four-dimensional spacetime or *Minkowski space*. They are labeled by $x^\mu = (ct, x, y, z)$, with ct taken as the zeroth coordinate. The set of coordinates is also called the position 4-vector. We wish to find a transformation between such coordinates in two systems, with the linear form

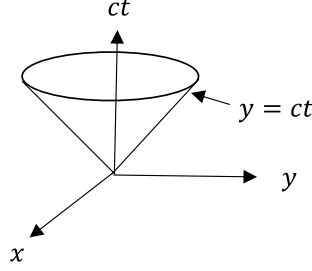


Fig. 2.1 The light cone in two space and one time dimension

$$x'^{\mu} = \sum_0^3 a^{\mu}_{\nu} x^{\nu} = a^{\mu}_{\nu} x^{\nu}. \quad (2.2)$$

Notice that in (2.2) we simply omitted the summation sign with the understanding that repeated indices are to be summed over. This is the famous *Einstein summation convention* which we will use henceforth; it makes the equations look much simpler.

The light cone equation (2.1) may be written in matrix notation as.

$$(ct, x, y, z) \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix} = 0 \quad (2.3a)$$

It may also be written using summation indices, called tensor component notation, as

$$x^{\mu} g_{\mu\nu} x^{\nu} = 0 \quad (2.3b)$$

The array $g_{\mu\nu}$ defined in (2.3a) is called the *Lorentz metric*. Notice that the order in which we write factors in (2.3b) is unimportant (see Exercise 2.2). In order that the equation of the light cone be invariant we now demand that the quantity $s^2 = x^{\mu} g_{\mu\nu} x^{\nu}$ be unchanged under the coordinate transformation (2.2); if it is zero in one frame it is zero in all frames related by the transformation (2.2). Thus we write in the original system and in the primed system,

$$\begin{aligned} s'^2 &= x'^{\mu} g_{\mu\nu} x'^{\nu} = (a^{\mu}_{\alpha} x^{\alpha}) g_{\mu\nu} (a^{\nu}_{\beta} x^{\beta}) = x^{\alpha} (a^{\mu}_{\alpha} g_{\mu\nu} a^{\nu}_{\beta}) x^{\beta}, \\ s^2 &= x^{\alpha} g_{\alpha\beta} x^{\beta}, \end{aligned} \quad (2.4)$$

and set them equal. Since the coordinates label an arbitrary event or point in the 4-space we find the following relation for the transformation.

$$g_{\alpha\beta} = a^{\mu}_{\alpha} g_{\mu\nu} a^{\nu}_{\beta} \quad (2.5a)$$

This relation (2.5a) defines the *Lorentz group* of transformations. The quantity s^2 plays the role of a four-dimensional distance or arc length. We thus say that the 4-distance is invariant under transformations in the Lorentz group. In matrix form the defining relation (2.5a) may be expressed as

$$G = A^T G A, \quad (2.5b)$$

where the T denotes the transpose matrix; you are asked to verify this in Exercise 2.3.

Example 2.1 Here are some examples of transformations in the Lorentz group. For relative motion at velocity v in the x direction there is the Lorentz transformation (1.18) that we studied in Chap. 1, which we repeat here with all four coordinates displayed,

$$A = \begin{pmatrix} \gamma & -\beta\gamma & 0 & 0 \\ -\beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \beta = v/c, \quad \gamma = 1/\sqrt{1 - \beta^2}. \quad (2.6)$$

Rotation about the z axis by angle θ is also a Lorentz transformation,

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos \theta & \sin \theta & 0 \\ 0 & -\sin \theta & \cos \theta & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \quad (2.7)$$

You should show that these are indeed in the Lorentz group as defined in (2.5a), and as requested in Exercise 2.5.

2.2 Four-Vectors and Tensors

We have called the set of coordinates of an event in spacetime the position 4-vector; the position 4-vector is the archetype of a contravariant 4-vector, which we now define in general as any set of 4 quantities which transform under a Lorentz transformation as

$$\bar{V}^\alpha = a^\alpha_\tau V^\tau. \quad (2.8)$$

That is, a *contravariant 4-vector* is a set of quantities that transforms like the coordinates. We will often refer to a contravariant 4-vector as simply a 4-vector.

We define another 4-component object with a lower index using the Lorentz metric,

$$V_\alpha = g_{\mu\nu} V^\nu, \quad (2.9)$$

which we call a *covariant 4-vector*. For example, the covariant position 4-vector is.

$$x_\mu = (ct, -x, -y, -z). \quad (2.10)$$

The operation in (2.10) is called lowering an index. An index may be raised similarly with the inverse of the Lorentz metric, which we denote as $g^{\mu\nu}$,

$$V^\alpha = g^{\alpha\nu} V_\nu, \quad g^{\alpha\lambda} g_{\lambda\omega} = \delta^\alpha_\omega. \quad (2.11)$$

You may easily verify that (2.10) and (2.11) are consistent. From the specific form of the Lorentz metric it is easy to see that the inverse of the Lorentz metric is simply the Lorentz metric itself, which is a convenient fact,

$$g^{\alpha\lambda} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}. \quad (2.12)$$

Since the two arrays in (2.12) and (2.3a) are the same the difference in index position is at this point purely for notational convenience. This will not be true later in a more general context. We also define for convenience a mixed index object, denoted by

$$g^\tau_\alpha = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \delta^\tau_\alpha. \quad (2.13)$$

This is called the Kronecker delta, equivalent to the identity matrix. Since it is symmetric the index order is irrelevant.

Let us now ask how covariant vectors transform as we go to a new coordinate system, labeled with a bar. We find from above that in the new system.

$$\bar{V}_\alpha = g_{\alpha\tau} \bar{V}^\tau = g_{\alpha\tau} (a^\tau_\beta V^\beta) = g_{\alpha\tau} (a^\tau_\beta g^{\beta\lambda} V_\lambda) = (g_{\alpha\tau} a^\tau_\beta g^{\beta\lambda}) V_\lambda. \quad (2.14)$$

We therefore define a new array called $b_\alpha{}^\lambda$ and rewrite (2.14) as.

$$\bar{V}_\alpha = b_\alpha{}^\lambda V_\lambda, \quad b_\alpha{}^\lambda \equiv g_{\alpha\tau} a^\tau_\beta g^{\beta\lambda}. \quad (2.15)$$

We call any quantity that transforms as in (2.15) a covariant 4-vector; it is consistent with the definition in (2.9). Note how similar the transformation law is to that for a contravariant 4-vector in (2.8). Note also that the index positions in the transformation matrices are relevant in (2.8) and (2.15).

Example 2.2 For the elementary Lorentz transformation in (2.6) we may calculate the array $b_\alpha{}^\lambda$ to be.

$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \gamma & -\beta\gamma \\ -\beta\gamma & \gamma \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} \gamma & \beta\gamma \\ \beta\gamma & \gamma \end{pmatrix}. \quad (2.16)$$

Here we have again suppressed the irrelevant y and z coordinates.

There is an important orthogonality relation between the transformation arrays $a^\alpha{}_\tau$ and $b_\alpha{}^\lambda$ that follows from the definition (2.15). From the definition of the Lorentz group in (2.5a), and using (2.11) and (2.15) we obtain.

$$a^\mu{}_\alpha b_\mu{}^\tau = a^\mu{}_\alpha (g_{\mu\nu} a^\nu{}_\beta g^{\beta\tau}) = (a^\mu{}_\alpha g_{\mu\nu} a^\nu{}_\beta) g^{\beta\tau} = g_{\alpha\beta} g^{\beta\tau} = \delta_\alpha^\tau \quad (2.17)$$

In matrix notation we may express this as

$$A^T B = I, \quad B = (A^T)^{-1} \quad (2.18)$$

From this it is also easy to see that.

$$a^\alpha{}_\omega b_\lambda{}^\omega = \delta^\alpha{}_\lambda \quad (2.19)$$

Equations (2.17) and (2.19) will be very useful, and provide a preview of how similar things work in the general theory. We may also now give an elegant alternative form for the Lorentz group definition, using (2.5a) and (2.19),

$$g_{\mu\nu} = b_\mu{}^\lambda b_\nu{}^\omega g_{\lambda\omega}. \quad (2.20)$$

We will return to this shortly and interpret its meaning and see why it is elegant.

Example 2.3 We may show that the 4-vector inner product $V^\mu V_\mu$ is invariant using the transformation properties and the orthogonality relation (2.17).

$$\overline{V}^\mu \overline{V}_\mu = (a^\mu{}_\alpha V^\alpha) (b_\mu{}^\beta V_\beta) = V^\alpha (a^\mu{}_\alpha b_\mu{}^\beta) V_\beta = V^\alpha \delta_\alpha^\beta V_\beta = V^\alpha V_\alpha. \quad (2.21)$$

Such invariant quantities are of great importance throughout relativity theory.

The above vectors and the metric are all examples of tensors. We define a general tensor by its transformation properties when going to a new coordinate system,

$$\overline{T}^{\gamma\delta\ldots}_{\kappa\rho\ldots} = a^\gamma_c a^\delta_d \ldots b_\kappa^n b_\rho^r \ldots T^{cd\ldots}_{nr\ldots} \quad (2.22)$$

We call this a tensor contravariant in the upper indices, and covariant in the lower indices. Thus V^μ is a contravariant tensor of size or rank 1, V_μ is a covariant tensor of rank 1, $g_{\mu\nu}$ is a covariant tensor of rank 2 as seen from (2.20), and so on for any rank.

An alternative definition of the Lorentz group can now be given: under a Lorentz transformation the Lorentz metric transforms as a covariant tensor of rank 2 and also remains the same! That is it is invariant

$$\overline{g}_{\mu\nu} = b_\mu^\lambda b_\nu^\omega g_{\lambda\omega} \quad (2.23)$$

In the more general theory of tensors in any coordinate system most of the above relations have natural generalizations, and in many cases the mathematics and notation make the general theory more transparent, as we will show in Part II.

Exercises

- 2.1 The Galilean transformation is one example of a linear transformation (2.1); what is the matrix a^μ_ν for it? Check that the equation of the light cone is not invariant under a Galilean transformation.
- 2.2 Convince yourself that the order used in a tensor equation like (2.3b) is irrelevant, or $x^\alpha g_{\alpha\beta} x^\beta = g_{\alpha\beta} x^\alpha x^\beta = x^\alpha x^\beta g_{\alpha\beta}$. This is because the elements of the arrays are simply numbers. The arbitrary order is a nice feature of the tensor notation.
- 2.3 Denote the matrix of the transformation coefficients by A and the matrix of the Lorentz metric by G , and show that the defining relation for the Lorentz group (2.5a) may be written in matrix notation as $G = A^T G A$.
- 2.4 Show that the Lorentz group as defined by (2.5a) is indeed a group according to the strict mathematical definition (you may want to review the definition of a group).
- 2.5 Verify that the transformations (2.6) and (2.7) are in the Lorentz group by verifying that they obey (2.5a). Find several more examples.
- 2.6 Show that a scalar, or invariant, times a 4-vector is a 4-vector. Is the difference between two 4-vectors a 4-vector? How about the derivative of a 4-vector with respect to a scalar parameter?
- 2.7 Show from the orthogonality properties of the transformations in (2.19) that the tensor inner product $T^{\alpha\beta}_\sigma S^\sigma_{\alpha\beta}$ is invariant. This generalizes Example 2.3.

Chapter 3

The Motion of Particles



Abstract This chapter deals with the motion of particles, their energy and momentum and acceleration, and emphasizes the geometric view of motion in space-time. In particular it demonstrates that special relativity is not limited to motion at constant velocity.

3.1 Energy and Momentum

The previous chapter contained a lot of formalism and little discussion of the physical world. Now it is time to see that the formalism we have developed can make physics more clear and easier (Schwartz 1968; Taylor 1963). We will consider some examples of 4-vectors in physics. As in classical mechanics we first consider the trajectory of a particle. Its position can be described by giving the functions of time $x(t)$, $y(t)$, $z(t)$ in some inertial lab frame; we thereby have the position 4-vector $(ct, x(t), y(t), z(t))$ as a function of time in that frame. The trajectory is a curve in four-dimensional spacetime and is also called the world-line of the particle. We illustrate it for two space dimensions in Fig. 3.1. Since the particle moves at less than the velocity of light the trajectory lies inside a light cone with vertex on any point of the trajectory, called the *local light cone*.

First consider an inertial coordinate system centered on a uniformly moving particle; recall that it is called the proper or rest frame of the particle. In this frame the position 4-vector is $x^\mu = (c\tau, 0, 0, 0)$, where τ is the time that a clock attached to the particle would measure, which we call the proper time. However, since $x^\mu x_\mu$ is an invariant we may write a relation that gives the proper time in *any* frame

$$c^2 \tau^2 = x^\mu x_\mu = c^2 t^2 - \vec{x}^2. \quad (3.1)$$

We emphasize that the proper time is an invariant, as is obvious from this expression!

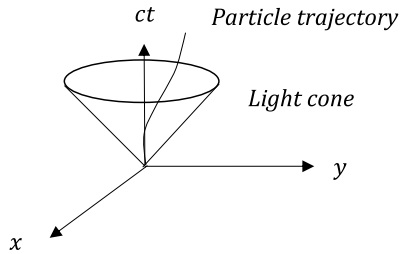


Fig. 3.1 The trajectory or world line of a moving particle

Next consider the trajectory of a particle which does not move uniformly but may accelerate and change velocity. For the trajectory of such a particle we consider short intervals of space and time along the trajectory. The differential of the 4-vector position, dx^μ , is also a 4-vector (as we noted in Exercise 2.6) so we may define an invariant proper time interval along the trajectory, in analogy with the above, as

$$c^2 d\tau^2 = c^2 dt^2 - d\vec{x}^2 = ds^2 \quad (3.2)$$

The quantity $ds^2 = c^2 d\tau^2$ in (3.2) is referred to as the line element; (3.2) is the differential analog of (3.1). From it we can obtain a useful relation between the lab time interval dt and the corresponding proper time interval $d\tau$. From (3.2) we may write

$$c^2 (d\tau/dt)^2 = c^2 - (d\vec{x}/dt)^2 = c^2 - v^2, \quad (3.3)$$

where v is the instantaneous velocity of the particle. Solving this for $dt/d\tau$ we find

$$\frac{dt}{d\tau} = \frac{1}{\sqrt{1 - v^2/c^2}} = \frac{1}{\sqrt{1 - \beta^2}} = \gamma. \quad (3.4)$$

This agrees with the time dilation relation (1.20), which we obtained in Chap. 1, but now applied to time intervals along the trajectory of a nonuniformly moving particle.

Having discussed the position 4-vector let us use it to construct some other 4-vectors which are useful in physics. Clearly we can consider the path of a particle as a function of the invariant proper time τ , that is $x^\mu(\tau)$; this has many advantages over using t as the independent variable. For example, the derivative of $x^\mu(\tau)$ with respect to τ is a 4-vector, which we will call the 4-velocity. We may write it explicitly as

$$u^\beta = \frac{dx^\beta}{d\tau} = \left(c \frac{dt}{d\tau}, \frac{d\vec{x}}{d\tau} \right) = \frac{dt}{d\tau} \left(c, \frac{d\vec{x}}{dt} \right). \quad (3.5)$$

Using the γ factor in (3.4) we may put this in simple and elegant form

$$u^\beta = \gamma(c, \vec{v}). \quad (3.6)$$

We emphasize that this is defined for any particle moving at velocity v , and not only for uniformly moving particles. In the instantaneous proper frame, where $v = 0$ and $\gamma = 1$, the square of the 4-velocity is obviously c^2 ; since it is an invariant it is thus equal to c^2 in any frame.

A most important 4-vector is the 4-momentum, which we construct from the velocity 4-vector in the same way as we construct the 3-vector momentum in classical mechanics, that is as the product of mass and velocity,

$$p^\mu = mu^\mu. \quad (3.7)$$

For low velocities the space components are approximately equal to the classical momenta, since γ approaches 1 at low velocities,

$$m\gamma\vec{v} = m\vec{v} + O(v^3/c^3). \quad (3.8)$$

The zeroth component times c is approximately the classical kinetic energy plus mc^2 ,

$$m\gamma c^2 = mc^2 + \frac{mv^2}{2} + O(v^4/c^2). \quad (3.9)$$

For this reason the 4-momentum is also called the energy-momentum vector. This 4-vector is the true momentum of relativistic physics; we identify the zeroth component as the relativistic energy divided by c and the space part as the relativistic momentum (Einstein 1923). That is

$$E = m\gamma c^2, \quad p^i = m\gamma v^i, \quad (3.10a)$$

$$p^\beta = (E/c, p^i). \quad (3.10b)$$

Note how this new definition of energy and momentum is forced on us by the formalism of special relativity when 4-vectors are viewed as the basic quantities.

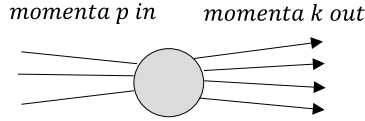


Fig. 3.2 The interaction of particles illustrates conservation of 4-momentum

Let us see if this definition of energy and momentum makes physical sense. If 4-momentum is conserved in an interaction in one reference frame, as in Fig. 3.2, we may express the fact as

$$p^\mu(\text{total in}) = \sum_i p_i^\mu = k^\mu(\text{total out}) = \sum_j k_j^\mu. \quad (3.11)$$

That is the total energy and momentum in are equal to the total energy and momentum out. Since both sides are 4-vectors they transform the same way in going to another reference frame, and the same equation holds; that is energy and momentum are conserved in the other frame. It is moreover nice that both energy and momentum conservation are contained in a single 4-vector equation. This is what we mean when we say that an equation or a law is *covariant* or *form invariant*: it has the same form in any reference frame. Clearly it is reasonable to expect that the fundamental laws of nature should be covariant, and in relativity this is indeed a basic postulate.

Example 3.1 Let us evaluate the invariant $p^\sigma p_\sigma$. This is most easily done in the proper frame of the particle, where we know from (3.6) that the 4-velocity is $u^\mu = (c, 0, 0, 0)$; hence

$$p^\sigma p_\sigma = m^2 u^\sigma u_\sigma = m^2 c^2. \quad (3.12)$$

This is also obvious since the square of the 4-velocity is c^2 as we noted after (3.6).

Example 3.2 There is always a reference frame in which a system of particles has a net momentum of zero, naturally called the center of momentum frame. To see that it exists first choose any inertial frame, and calculate the total energy E and momentum $\vec{P}$ of the particles. Then orient the x axis along the momentum so the momentum 4-vector is $P^\mu = (E/c, P, 0, 0)$. In a prime frame moving at v along that x axis the energy and momentum are, from the Lorentz transformation (1.18),

$$\frac{E'}{c} = \gamma \left(\frac{E}{c} \right) - \beta \gamma P, \quad P' = -\beta \gamma \left(\frac{E}{c} \right) + \gamma P, \quad (3.13)$$

If we choose $\beta = Pc/E$ we see that the momentum is zero in the primed frame. The fact that the center of momentum velocity is $\beta = Pc/E$ is often useful in relativistic kinematics.

Recall that in classical mechanics the relation between kinetic energy and momentum of a particle is given by $E = p^2/2m$. There is a very useful analog of this in special relativity. The square of the 4-vector momentum can be expressed in two ways: first it is an invariant which we calculated in Example 3.1 to be m^2c^2 ; second, it is the square of the momentum 4-vector, $(E/c)^2 - \vec{p}^2$. We thus obtain the energy as a simple function of the momentum,

$$E^2 = (mc^2)^2 + (\vec{p}c)^2. \quad (3.14)$$

This is very useful in doing kinematics problems. We may also define the kinetic energy as the relativistic energy in (3.14) minus the rest energy.

3.2 Acceleration

In the simple approach to special relativity in Chap. 1 we studied the Lorentz transformation between *uniformly* moving systems; this in no way restricts special relativity to uniform motion, and accelerated motion fits nicely into the conceptual and mathematical framework. We first define the 4-vector acceleration of a particle in the obvious way, as the derivative of the 4-vector velocity with respect to the proper time of the particle,

$$a^\mu = \frac{du^\mu}{d\tau} = \frac{d^2x^\mu}{d\tau^2}. \quad (3.15)$$

We may express this in terms of the classical velocity and acceleration, which involve t derivatives, not τ derivatives. To do this we use the expression for the 4-velocity in (3.6) and the relation between $d\tau$ and dt in (3.4), which implies $d/d\tau = \gamma(d/dt)$, to obtain

$$a^\mu = \frac{du^\mu}{d\tau} = \frac{d}{d\tau}(\gamma c, \gamma \vec{v}) = \gamma \frac{d}{dt}(\gamma c, \gamma \vec{v}) = \left(\gamma c \frac{d\gamma}{dt}, \gamma^2 \frac{d\vec{v}}{dt} + \gamma \vec{v} \frac{d\gamma}{dt} \right). \quad (3.16)$$

The derivative of the velocity is of course the classical acceleration $\vec{a} = d\vec{v}/dt$, while the derivative of γ is easy to calculate as

$$\gamma \frac{d\gamma}{dt} = \frac{1}{2} \frac{d\gamma^2}{dt} = \frac{1}{2} \left(1 - \frac{v^2}{c^2} \right)^{-2} \frac{d}{dt} \left(\frac{v^2}{c^2} \right) = \left(\frac{\gamma^4}{c^2} \right) \vec{v} \cdot \frac{d\vec{v}}{dt} = \left(\frac{\gamma^4}{c^2} \right) \vec{v} \cdot \vec{a}. \quad (3.17)$$

Thus

$$a^\mu = \left(\frac{\gamma^4}{c} (\vec{v} \cdot \vec{a}), \vec{v} \frac{\gamma^4}{c^2} (\vec{v} \cdot \vec{a}) + \gamma^2 \vec{a} \right). \quad (3.18)$$

In particular, in the proper frame where the velocity vanishes instantaneously, we have

$$a^\mu = (0, \vec{a}), \text{ proper frame.} \quad (3.19)$$

This should not be surprising.

Example 3.3 From the above we can show that the 4-velocity and the 4-acceleration are orthogonal, that is the invariant $a^\beta u_\beta = 0$. One way to see this is to evaluate both the velocity and acceleration 4-vectors in the proper frame; in that frame the 4-velocity (3.6) has only a zeroth component while the acceleration (3.19) has no zeroth component, so the inner product is zero. Another way is to recall that the square of the 4-velocity is the constant c^2 , so that

$$\frac{d}{d\tau}(u^\beta u_\beta) = 0 = \frac{1}{2} u_\beta \frac{du^\beta}{d\tau} = \frac{1}{2} u_\beta a^\beta. \quad (3.20)$$

3.3 Accelerated Motion

Now we are ready to study the trajectory of an accelerated particle in one space dimension. We will think of the particle as a small rocket, since a rocket is built with internal means of acceleration. In doing this we will see how convenient the concept of rapidity is for such calculations (Misner 1973).

Consider a rocket moving in the x direction as in Fig. 3.3. The proper time τ provides a convenient parameter for defining the trajectory of the rocket, $ct(\tau)$, $x(\tau)$. At proper time τ the rocket has velocity v in the lab system S , while in its instantaneous rest frame S' its velocity is of course zero. A short time $d\tau$ later its velocity in S' is given in terms of the acceleration by

$$dv' = ad\tau, \quad d\beta' = (a/c)d\tau, \quad a = \text{proper acceleration.} \quad (3.21)$$

The proper acceleration is that measured in the proper frame, where the rocket is instantaneously at rest. In the lab frame S the rocket velocity after the little time interval and velocity change is gotten from the velocity addition relation in Exercise 1.3,

$$\beta(\text{after } d\tau) = (\beta + d\beta')/(1 + \beta d\beta'). \quad (3.22)$$

Thus the change in the lab velocity of the rocket to first order in $d\beta'$ is

$$d\beta = (1 - \beta^2)d\beta' = d\beta'/\gamma^2 = (a/c)d\tau/\gamma^2. \quad (3.23)$$

This is a differential relation giving β as a function of τ and the proper acceleration since γ is a function of β ; if the acceleration were given as a function of τ we could integrate (3.23) to get $\beta(\tau)$.

However there is a more elegant way to analyze (3.23) in terms of rapidity. From the definition of rapidity θ in (1.28) we have

$$\beta = \tanh \theta, \quad d\beta = \text{sech}^2 \theta d\theta = d\theta / \cosh^2 \theta = d\theta / \gamma^2,$$

$$d\theta = \gamma^2 d\beta. \quad (3.24)$$

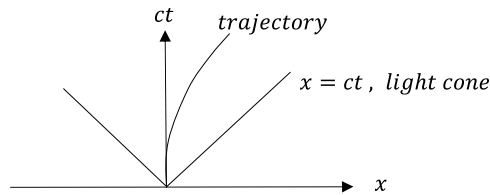


Fig. 3.3 Trajectory of the accelerated particle or rocket

Then from (3.23) above we get an elegant differential relation between rapidity and the proper acceleration.

$$\frac{d\theta}{d\tau} = \frac{a}{c}. \quad (3.25)$$

That is, the derivative of rapidity with respect to rocket proper time is the proper acceleration divided by c . Accordingly if we are given the acceleration as a function of proper time we may suppose that (3.25) has been solved to give the rapidity as a function of the proper time (see Exercise 3.3 for the case of constant acceleration).

Now we may easily integrate to get the spacetime trajectory of the rocket. From the definition of θ and the fundamental Lorentz relation $dt/d\tau = \gamma$ we can write β as

$$\beta = \frac{dx}{d(ct)} = \left(\frac{1}{c}\right) \frac{dx}{d\tau} \frac{d\tau}{dt} = \frac{1}{c\gamma} \frac{dx}{d\tau}. \quad (3.26)$$

We thereby get a differential relation between dx and $d\tau$

$$dx = \beta \gamma c d\tau = (\tanh \theta \cosh \theta) c d\tau = \sinh \theta c d\tau. \quad (3.27)$$

Similarly we can get a differential relation for $cd\tau$

$$cdt = c \frac{dt}{d\tau} d\tau = c\gamma d\tau = \cosh \theta c d\tau. \quad (3.28)$$

Using (3.27) and (3.28) we integrate to obtain the trajectory in terms of the rapidity, which we may take to be a known function of proper time

$$ct = \int_0^\tau \cosh \theta d\tau, \quad x = c \int_0^\tau \sinh \theta d\tau. \quad (3.29)$$

This is a complete solution to the general problem in one dimension; with the proper acceleration given as a function of proper time equation (3.25) gives $\theta(\tau)$ and (3.29) gives the trajectory. You should work out the special case of constant acceleration as requested in Exercise 3.3; the result is a hyperbolic trajectory.

3.4 Curves and Arc Lengths

The lengths of lines and curves in the spacetime of special relativity have some peculiar and interesting properties. Let us study the time elapsed for travelers aboard rocket ships having diverse trajectories, curves in spacetime, using the geometric view that we have developed. The proper time interval for such a traveler is equal to the square root of the line element divided by c , as in (3.2),

$$cd\tau = ds = \sqrt{c^2dt^2 - d\vec{x}^2} = cdt\sqrt{1 - (d\vec{x}/cdt)^2} = \sqrt{1 - \beta^2}cdt, \quad (3.30)$$

where the space and time intervals are measured in some inertial system such as our lab. Notice that this has meaning only so long as the velocity of the rocket ship is less than c , for otherwise the proper time becomes imaginary. That is the trajectory must always have a slope of over 45° . The time elapsed for a traveler is thus simply the integral of the line element along the trajectory, or the arc length of the curve between initial and final points in spacetime,

$$c\tau = s = \int_i^f \sqrt{1 - \beta^2}cdt. \quad (3.31)$$

It is obvious from the integrand in (3.31) that this arc length is largest when the velocity of the rocket remains small along the trajectory. In particular the longest curve in spacetime for the roundtrips shown in Fig. 3.4 is the straight line along the time axis; any other curve is shorter, and as the curve approaches the 45° lines (light cone) its length approaches zero!

A straight line of this type is the *longest* distance between 2 points in spacetime, whereas it is the *shortest* distance between two points in Euclidean space. The minus sign in the line element (3.2) produces this profoundly different behavior. A physical consequence of this is that someone who leaves earth and travels at high velocity, say to a nearby star, and returns to earth will be younger than indicated by an earthbound clock. The infamous “twin paradox” is based on this peculiar behavior. See Exercise 3.6.

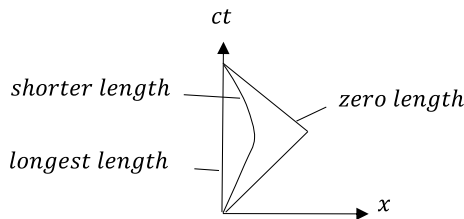


Fig. 3.4 The longest arc length (elapsed travel time) is the straight line along the time axis; the arc length along the 45° lines is zero

This is a good place to mention an arbitrary sign choice we have made in the last three chapters. The Lorentz metric as defined in (2.3a) contains a single plus sign and three minus signs. With that choice for the metric the relation between arc length and proper time for a particle trajectory is $c^2 d\tau^2 = ds^2$ so the proper time interval is $d\tau = ds/c$; it is positive for a moving particle. Some authors instead choose the opposite sign for the Lorentz metric since it contains only a single minus sign. With this choice the relation between proper time and arc length becomes the somewhat awkward $d\tau = \sqrt{-ds^2}/c$.

One drawback of our sign choice is that the Einstein equations that we will study in Part III contain a minus sign between the left side describing geometry and right side describing energy and momentum.

Another way to view the choice of signs is that we might want to think of time as more “important” than space, or space as more “important” than time; the choice is obviously one of taste and notational convenience. It is also relevant that during much of the twentieth century the choice we have made was the dominant one on the west coast of the US and the other was the dominant one on the east coast! At present both choices are common; the text of Misner Thorne and Wheeler contains a large table of sign conventions for the metric and other tensors used in relativity theory until 1973 (Misner 1973). In particle physics the choice we have made is prevalent (Bjorken 1963; Griffiths 1987). See Exercise 3.7.

Exercises

- 3.1. Consider a particle of mass M that decays at rest and turns into two particles of equal mass m , with $2m < M$. What is the energy of each decay particle? What is the momentum of each? What is the kinetic energy of each, that is the energy minus the rest energy mc^2 ?
- 3.2. For a particle of zero rest mass such as a photon the relations for energy and momentum in (3.10a) are not meaningful, but the relation between energy and momentum in (3.14) remains reasonable. Using elementary quantum theory and the Planck and de Broglie relations for energy and momentum show that (3.14) is indeed correct for a photon, that is $E^2 = \vec{p}^2 c^4$.
- 3.3. Take the case of constant proper acceleration, $a = \text{constant}$, and solve for the trajectory from (3.25) and (3.29).
- 3.4. Show that the trajectory is a hyperbola in the variables ct and x , and the asymptotic velocity is c .
- 3.5. Draw a nice graph of the hyperbolic motion from Exercise 3.4, and from it show that a photon sent from $x = 0$ after time $t = c/a$ will never catch the rocket.
- 3.6. There have been many papers and books written on the twin paradox, wherein a twin who travels at high velocity to a nearby star system and returns is younger than his twin who remains on earth. Think about this and convince yourself that there is no contradiction. The problem is discussed in many reputable books, for example the readable Feynman lectures (Feynman 1963; Schutz 2009). However it is also a favorite topic in books and articles by people with limited or incorrect understanding of relativity, so beware.

- 3.7. Work through the development of energy and momentum theory in Sect. [3.1](#) using the opposite sign for the Lorentz metric as discussed at the end of this chapter. What is your own preference for the sign choice of the metric tensor in special relativity?

Part II

Vectors and Tensors

We now begin to develop the mathematics needed for general relativity. General relativity was invented and developed by Einstein and others using the classic tensor index calculus invented by nineteenth-century mathematicians such as Riemann and Ricci and Levi-Cevita. In this approach, vectors are viewed and treated as n -tuples just as we treated them in special relativity in Part I; the metric tensor is treated as an n by n array and so forth. We call this the classic or index or component view of tensors. Most physics applications are still done using the component view. It is convenient that the component view relies on elementary vector and matrix theory familiar to all physicists. One important feature of the component view is that vectors and tensors are fundamentally tied to a coordinate system.

An alternative view was developed later in the twentieth century, and now favored by many mathematicians and theorists, which we may call the intrinsic or invariant abstract view. In this view, vectors and tensors and forms are invariant abstract objects independent of coordinate systems. As an example, one can think of an abstract 3-vector in the usual way as an arrow in 3-space. The most important feature of the abstract view is its independence of a coordinate system; some thus consider it more physical. The component and abstract views are related simply in that the tensor components arise as coefficient arrays when the abstract tensor is expanded in a basis. As such there is a one-to-one correspondence between almost all the concepts and theorems in the component and abstract views.

The relation between the abstract and component views is somewhat like the relation of classic Greek geometry, using abstract idealized points and lines and curves, to Cartesian analytic geometry using coordinates and n -tuples. We will discuss most topics first from the component view and then from the abstract view (Bergmann 1942; Rindler 1969, Weinberg 1972; Adler 1975; Kenyon 1990)

Much of the mathematics in this part is a fairly easy generalization of the vector calculus used in classical mechanics and electromagnetism, and the 4-vector ideas of special relativity. Central concepts are those of a Riemann space, vectors and tensors and forms, affine connections, geodesics that generalize the straight lines of elementary geometry, and covariant derivatives which generalize the derivatives of elementary calculus (Lawrie 1990; Arfken 1970). Most of this part is mathematical,

but at the end of Chap. 5 we will study classical mechanics in light of some of the geometrical concepts.

One important mathematical subject that we will only treat later in Part III is curvature and the Riemann tensor; curvature is intimately tied to the geometric interpretation of gravity.

Chapter 4

Riemann Spaces and Tensors



Abstract In this chapter we begin to study the mathematics needed for general relativity. Quite general spaces such as we will use to describe spacetime and gravity were first developed by nineteenth century mathematicians such as Gauss and Riemann. The most important mathematical objects in such spaces are vectors and tensors. We will treat these using both the classic index notation and a more modern abstract notation.

4.1 Riemann Spaces

We now make the transition from the Minkowski spacetime of special relativity to more general spaces and coordinate systems and the mathematical objects in them, vectors and tensors and forms. There are standard reference texts dating over many years (Pauli 1958; Bergmann 1942; Rindler 1969; Weinberg 1972; Misner 1973; Adler 1975; Kenyon 1990). Here we will rely heavily on examples to illustrate the basic ideas.

Think of a physical space such as the surface of a blackboard or sphere or torus as in Fig. 4.1, or the Euclidean 3-space of classical physics. In particular include the spacetime of special relativity that we studied in Part I. We imagine a *marker* system or *labeling* system or *coordinate* system to specify the points in the space with a set of real numbers. In general there will be many ways to set up such a marker system, and we assume that there will be transformations between them. An excellent example to remember is the Euclidean 3-space of classical geometry and physics, labeled by Cartesian or spherical coordinates. We denote the transformation between two coordinate systems, denote them as unprimed and primed, by a set of functions

$$x'^j = f^j(x^n). \quad (4.1)$$

The functions f^j are assumed to be continuous monotonic one-to-one and differentiable as often as needed. The transformation therefore has an inverse. For brevity we usually denote the transformation and its inverse in shorthand notation,

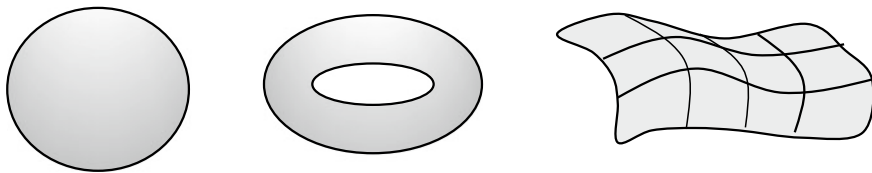


Fig. 4.1 Some 2-dimensional spaces: a sphere, a torus, and an odd shaped 2-surface with coordinates lines shown

$$x'^j = x'^j(x^n), \quad x^k = x^k(x'^j). \quad (4.2)$$

The square array of derivatives we denote as

$$\frac{\partial x'^j}{\partial x^n}, \quad \frac{\partial x^j}{\partial x'^k}. \quad (4.3)$$

These are the transformation or Jacobian matrices familiar from elementary calculus.

Loosely speaking a space coordinatized in several different ways by n real numbers as we use here is called an n -dimensional manifold. A manifold is defined as a space which locally resembles a Euclidean space and in which we can perform the usual analytic operations as in Euclidean space. Thus we can for example set up systems of differential equations in a manifold. See Appendix 1 for a more detailed discussion of the manifold idea.

The spaces of interest in physics usually have a well-defined distance between any two points. We therefore assume that between two nearby points in the space, separated by small coordinate distances dx^μ , there is a distance with physical meaning, or *line element*, given by a quadratic form

$$ds^2 = \sum_{\mu\nu} g_{\mu\nu} dx^\mu dx^\nu = g_{\mu\nu} dx^\mu dx^\nu. \quad (4.4)$$

Here we use as usual the Einstein summation convention, wherein a repeated index or dummy index is to be summed over. The line element is a direct generalization of the Pythagorean theorem of Euclidean geometry on a differential scale. The array $g_{\mu\nu}$ is called the *metric* or *metric tensor*; the Lorentz metric of special relativity is one important example. A space with such a distance measure or metric we will call a *Riemann metric space*. The phrase Riemannian manifold is more specific and often used, as elucidated in Appendix 1.

Since the expression (4.4) for the line element completely determines the metric tensor we will often refer to the line element as the metric.

The summation convention is very powerful in the sense that it simplifies the look of an equation; it is important to remember that repeated or dummy indices are summed over so they may be denoted by any convenient symbol. After a little practice the “index juggling” we will encounter in tensor equations becomes easy.

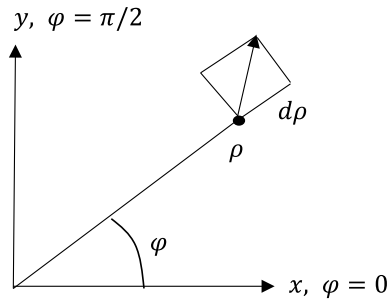


Fig. 4.2 Cartesian and polar coordinates for Euclidean 2-space, with the differential box labeled in polar coordinates

We will always assume the metric is symmetric; if it had an antisymmetric part then that part would not contribute to the line element, as you may verify in Exercises 4.1–4.3.

Example 4.1 A simple but nontrivial example of these ideas is Euclidean 2-space labeled with Cartesian or polar coordinates. Figure 4.2 shows the relation between the two and the differential box from which we may read off the line element.

The Cartesian coordinates as functions of the polar coordinates are

$$x = \rho \cos \varphi, \quad y = \rho \sin \varphi. \quad (4.5a)$$

and the polar coordinates as functions of the Cartesian coordinates are

$$\rho = \sqrt{x^2 + y^2}, \quad \varphi = \tan^{-1}(y/x). \quad (4.5b)$$

For the transformation from Cartesian coordinates (unbarred) to polar coordinates (barred) we differentiate and get the Jacobian matrix,

$$\frac{\partial \bar{x}^i}{\partial x^j} = \begin{pmatrix} \cos \varphi & \sin \varphi \\ -\sin \varphi / \rho & \cos \varphi / \rho \end{pmatrix} \downarrow i. \quad (4.6a)$$

For the transformation from polar to Cartesian coordinates we likewise get

$$\frac{\partial x^i}{\partial \bar{x}^j} = \begin{pmatrix} \cos \varphi & -\rho \sin \varphi \\ \sin \varphi & \rho \cos \varphi \end{pmatrix} \downarrow i. \quad (4.6b)$$

We have expressed both matrices in terms of the spherical coordinates; it is easy to switch to Cartesian coordinates if desired.

From the differential box in Fig. 4.2 it is easy to use the Pythagorean theorem to calculate the distance between nearby points, which gives the line element in the two coordinate systems

$$ds^2 = dx^2 + dy^2 \text{ Cartesian, } ds^2 = d\rho^2 + \rho^2 d\varphi^2 \text{ polar.} \quad (4.7)$$

Hence the metric tensor in the two systems is

$$g_{\mu\nu} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \text{ Cartesian, } g_{\mu\nu} = \begin{pmatrix} 1 & 0 \\ 0 & \rho^2 \end{pmatrix} \text{ polar.} \quad (4.8)$$

Since the coordinate lines are orthogonal in both systems the metric is diagonal, which is often convenient.

There is a fundamental difference between spaces like the Euclidean spaces alluded to in the above example, and the Minkowski space of special relativity. The Euclidean line element has all positive terms, but the Minkowski line element has one positive and three negative terms, which leads to many interesting physical effects as discussed in Chaps. 1–3. There is a theorem from classical matrix theory that allows us to categorize this property of a space in an interesting and useful way (Perlis 1952).

Signature Theorem *Consider a single point P in a metric space. One may find a coordinate system at P in which the metric tensor is diagonal and has $+1$ or -1 or 0 as diagonal elements. This form of the metric is called the Cayley-Sylvester canonical form, and the set of diagonal elements is called the signature; the signature is a unique and invariant characteristic of the metric at P . Moreover, the special coordinate system can be obtained by a linear transformation at P beginning with any coordinate system. We prove this theorem for two dimensions in Appendix 2.*

Thus the signature of Euclidean n -space is $(1, \dots, 1)$, that of Minkowski spacetime is $(1, -1, -1, -1)$ and so forth. In many of our 2-surface examples the signature will be $(1, 1)$, and in general relativity the signature will be the same as in special relativity $(1, -1, -1, -1)$. We will usually suppose the signatures of the spaces we study generally have no zeros, for a zero would imply that the metric determinant is zero and the metric has no inverse, which is a problematic situation as we will see.

Note an important point: the theorem says that there is a coordinate system where the metric has this special form at *any single given point*; in general one cannot find a coordinate system where the metric has this form throughout space or even in a small neighborhood. We will indeed show later that such a global system can be found only for a *flat* space, a term which we will later define more precisely.

Note that the overall sign of the metric tensor is arbitrary as we have discussed in Chap. 3. Other authors use the negative of our choice, so their signature is $(-1, 1, 1, 1)$. Both sign conventions have virtues and drawbacks but of course do not affect the physics (Misner 1973).

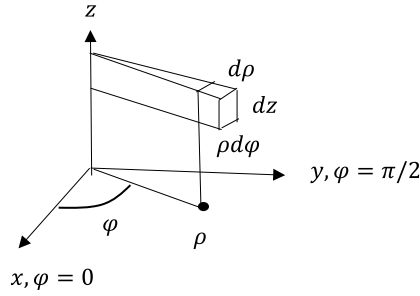


Fig. 4.3 Cylindrical coordinates in Euclidean 3-space. The differential box sides are $d\rho$ and $\rho d\varphi$ and dz

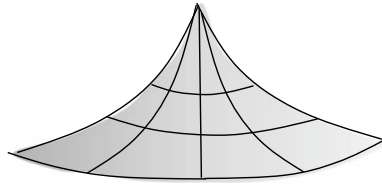


Fig. 4.4 The cylindrically symmetric curved 2-surface described by (4.11)

Example 4.2 Many of the basic ideas of Riemann spaces are well illustrated with curved 2-surfaces. To illustrate a few of these first consider Euclidean 3-space with cylindrical coordinates. Figure 4.3 shows the relation to Cartesian coordinates; the differential box gives the line element, much as we discussed in Example 4.1.

Pythagoras and Fig. 4.3 tell us that the line element is

$$ds^2 = d\rho^2 + \rho^2 d\varphi^2 + dz^2 \text{ cylindrical.} \quad (4.9)$$

A flat surface is defined by the equation $z = \text{constant}$, which gives a 2-dimensional surface with polar coordinates as in the above Example 4.1. A more general type of 2-surface results if we take z to be a function of ρ , $z = f(\rho)$. Then the surface line element becomes

$$\begin{aligned} ds^2 &= d\rho^2 + \rho^2 d\varphi^2 + f'^2 d\rho^2 \\ &= (1 + f'^2) d\rho^2 + \rho^2 d\varphi^2, \quad f' \equiv df/d\rho. \end{aligned} \quad (4.10)$$

This describes a curved cylindrically symmetric 2-surface. For example take the function f to be a decreasing exponential, so the 2-surface has the shape of a mountain peaked at the origin as shown in Fig. 4.4. More explicitly,

$$z = ae^{-\rho/a}, \quad a = \text{const.}, \quad ds^2 = (1 + e^{-2\rho/a})d\rho^2 + \rho^2 d\varphi^2. \quad (4.11)$$

This is one example of a curved 2-surface as a hypersurface in Euclidean 3-space.

Note also that the surface is not smooth at the origin.

Curvature is a clear intuitive idea for 2-surfaces like those in the above example. But note the important fact that not all 2-surfaces which we can define and study may be considered to be hypersurfaces in a Euclidean 3-space. Accordingly we need a more precise and general definition of curvature, also applicable to any number of dimensions, as we will discuss later in Chap. 8.

4.2 Vectors, Component View

We will first discuss mathematical objects in Riemann space from the point of view of their components, the view mainly used in the early part of the twentieth century for the invention and the early development of general relativity by Einstein and others (Pauli 1958; Bergmann 1942; Rindler 1969; Adler 1975). This is the approach we used in special relativity in Part I for Minkowski spacetime but now applied to a Riemann space. In Sect. 4.3, we will relate the component view to the more modern invariant abstract view, which became popular and fashionable in the later twentieth century (Misner 1973; Schutz 2009).

As we have already noted the component view may be termed the classic view, and is most useful for calculations such as finding solutions to the Einstein field equations and solving for the trajectories of moving objects. The abstract view can give a different perspective on the mathematics. The reader interested only in applications such as cosmology might choose to skim or skip the sections on the abstract view but could benefit from being exposed to both views.

The line element in (4.4) is the archetype of an invariant, a crucially important mathematical object; it is postulated to be the same in all coordinate systems. That is an invariant or *scalar* is defined as any quantity which has the same value in all coordinate systems, for example for an unprimed and a primed system,

$$\phi' = \phi \text{ scalar or invariant.} \quad (4.12)$$

The concept of an invariant is one of the most fundamental in relativity and all of physics. Virtually everything that theory predicts should be expressed as an invariant for comparing with experimental measurement since nature does not know or care about our choice of coordinates.

Note that in this chapter we generally will not limit ourselves to any specific number of dimensions or metric signature, and the indices that we will use may be either Latin or Greek.

The archetype of a vector, our next mathematical object, is the set of coordinate differentials dx^n along some given curve; the transformation law is easily calculated using the chain rule

$$x'^j = x'^j(x^n), \quad dx'^j = \frac{\partial x'^j}{\partial x^n} dx^n. \quad (4.13)$$

Any n -tuple which transforms according to (4.13) is termed a *contravariant* vector, for reasons we will discuss below,

$$V'^j = \frac{\partial x'^j}{\partial x^n} V^n \quad \text{contravariant vector components.} \quad (4.14)$$

We emphasize that according to this definition the coordinates x^n do *not* form a vector, unlike the situation in special relativity.

Another type of vector has as an archetype the gradient of a scalar, $\phi_{,j}$ where the comma denotes an ordinary derivative. This transforms by the chain rule according to

$$\frac{\partial \phi'}{\partial x'^k} = \frac{\partial x^j}{\partial x'^k} \frac{\partial \phi}{\partial x^j}, \quad \text{or} \quad \phi'_{,k} = \frac{\partial x^j}{\partial x'^k} \phi_{,j}. \quad (4.15)$$

Any n -tuple which transforms like the gradient in (4.15) is termed a *covariant* vector, a name we will justify below,

$$W'_k = \frac{\partial x^j}{\partial x'^k} W_j \quad \text{covariant vector components.} \quad (4.16)$$

Note carefully the position of the indices and primes in (4.14) and (4.16).

From the above definitions many simple but important theorems follow. Let us prove two of them that are relevant to vectors.

Theorem 1 *The Jacobian matrices of the transformation and the inverse transformation, in (4.3), are inverses of each other. To see this note that, by definition, the function of the inverse function is the identity function; that is we may write*

$$\bar{x}^j(x^k(\bar{x}^n)) = \delta_n^j \bar{x}^n. \quad (4.17)$$

From this we obtain, using the chain rule,

$$\frac{\partial \bar{x}^j}{\partial x^n} = \frac{\partial \bar{x}^j}{\partial x^m} \frac{\partial x^m}{\partial x^n} = \delta_n^j \quad \text{similarly} \quad \frac{\partial \bar{x}^j}{\partial x^k} \frac{\partial x^n}{\partial \bar{x}^j} = \delta_k^n, \quad (4.18)$$

which is the desired theorem. This is the generalization of the orthogonality relation on the Lorentz transformations of special relativity (2.17). Many other theorems follow from it.

Theorem 2 *The inner product of a contravariant and covariant vector, defined as $V^\beta W_\beta$, is a scalar. We use the transformation of the vectors and Theorem 1 to calculate the inner product in the barred system to see that it is an invariant,*

$$\bar{V}^\beta \bar{W}_\beta = \frac{\partial \bar{x}^\beta}{\partial x^\eta} \frac{\partial x^\sigma}{\partial \bar{x}^\beta} V^\eta W_\sigma = \delta_\eta^\sigma V^\eta W_\sigma = V^\eta W_\eta. \quad (4.19)$$

Thus we see that we may think of covariant vectors as objects that map contravariant vectors into scalars. We will return to this idea when we discuss forms in the next section.

4.3 Vectors and 1-Forms, Abstract View

We can connect the above idea of vectors as component n -tuples with the idea of intrinsic or abstract vectors, often represented in physics by arrows, and in the process introduce a definition of a metric. We may think of such vectors as intrinsic or abstract, but the word physical is also appropriate since they are taken to exist independently of the coordinate system and are invariant. These abstract vectors are taken to exist in an idealized physical world, whereas component vectors only exist when we represent them in terms of a specific coordinate system.

Look at a single point P in a Riemann space. We introduce a set of basis vectors $\vec{e}_k$ along the grid lines illustrated in Fig. 4.5 for two dimensions, but we do *not* assume the basis is orthonormal. The basis set spans a vector space associated with that point. Such a basis is naturally called a coordinate basis.

A small displacement $d\vec{s}$ along a curve or in some specified direction is given by

$$d\vec{s} = \vec{e}_j dx^j, \quad (4.20)$$

and its square is given by

$$\begin{aligned} d\vec{s}^2 &= ds^2 = (\vec{e}_j dx^j) \cdot (\vec{e}_k dx^k) = (\vec{e}_j \cdot \vec{e}_k) dx^j dx^k = g_{jk} dx^j dx^k, \\ g_{jk} &= \vec{e}_j \cdot \vec{e}_k. \end{aligned} \quad (4.21)$$

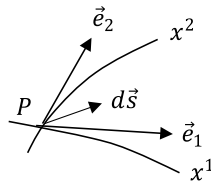


Fig. 4.5 A coordinate basis in two dimensions, with a small displacement vector $d\vec{s}$

The metric g_{jk} defined in this way, as the inner product of basis vectors, agrees with the line element expression (4.4) and implies that the metric is intrinsically symmetric. Note that, for example, a coordinate interval dx^1 along the first axis corresponds to an invariant physical distance

$$ds = \sqrt{g_{11}} dx^1, \quad (4.22)$$

Here g_{11} is the square of $\vec{e}_j$ and could be anything we choose; we assume in this example that g_{11} is positive. This shows clearly the role of the metric in relating coordinate distances to physical distances: only the combination of coordinates and metric has physical meaning. Note in particular that the dimensions of the coordinates need not be distances; for example they could be angles as in polar or spherical coordinates; but the dimension of the metric components must be such that the product in (4.22) is a distance.

Now consider any vector $\vec{V}$ at P . We expand it in the coordinate basis as we did for the small displacement vector in Fig. 4.5, and calculate its square

$$\begin{aligned} \vec{V} &= \vec{e}_j V^j, \\ \vec{V}^2 &= (\vec{e}_j V^j) \cdot (\vec{e}_k V^k) = (\vec{e}_j \cdot \vec{e}_k) V^j V^k = g_{jk} V^j V^k. \end{aligned} \quad (4.23)$$

Here V^k are the vector components with respect to the coordinate basis; they correspond to the contravariant component vector in (4.14).

The vector $\vec{V}$ can also be characterized in terms of its projections on the basis vectors; denoting these projections with lower indices we define explicitly,

$$V_i = \vec{V} \cdot \vec{e}_i = (\vec{e}_j V^j) \cdot \vec{e}_i = (\vec{e}_j \cdot \vec{e}_i) V^j = g_{ij} V^j. \quad (4.24)$$

The V_i correspond to the covariant component vector in (4.16). Just as in special relativity the metric lowers the index position according to (4.24). Conversely, we may define the inverse g^{jn} of the metric as the inverse of its associated matrix and invert the relation (4.24). This gives

$$g^{jn} g_{nk} = \delta_k^j, \quad V^j = g^{jk} V_k. \quad (4.25)$$

Thus we are led to the idea of lowering and raising indices with the metric just as in the previous Sect. 4.2. It is a natural extension of the ideas and notation of special relativity in Part I.

An important basic mathematical point which we emphasize here is that the real n -tuples V^i introduced with respect to a coordinate basis may be viewed in two separate ways:

1. As *components* of a vector with respect to a basis, as in (4.23).
2. As a *representation* of the vector.

This is a common situation in mathematics; for example relations in group theory may be represented in terms of matrices and row and column vectors. Probably the most important example is in quantum physics: the wave function may be viewed as an inner product of a Hilbert space state vector with a position eigenstate, or as a representation of the state.

Digression 4.1 Let us digress briefly to ask an important question: what happens if we use different coordinates and thus different basis vectors? The displacement $d\vec{s}$ represents a real physical distance, independent of the coordinate system, so it should not change, but its components change. Thus we may write for two systems, unprimed and primed,

$$d\vec{s} = \vec{e}_j dx^j = \vec{e}_j \left(\frac{\partial x^j}{\partial x'^k} \right) dx'^k = \vec{e}'_k dx'^k. \quad (4.26)$$

Since the coordinate displacements are arbitrary we must have the following transformation rule for the basis vectors and the differentials,

$$\vec{e}'_k = \left(\frac{\partial x^j}{\partial x'^k} \right) \vec{e}_j, \quad dx'^i = \left(\frac{\partial x'^i}{\partial x^k} \right) dx^k, \quad (4.27)$$

If we compare these two expressions we see that the coordinate differentials and the basis vectors transform in an *opposite* sense, what is called *contragrediently*. Objects which transform in the *same* sense are said to transform *cogradiently*.

Our next mathematical object is a 1-form: 1-forms comprise a dual space to vectors. They are defined to operate linearly on vectors to give real scalars; a 1-form $\tilde{p}$ operates on a vector $\vec{V}$ and maps it into a scalar according to the defining rule

$$\tilde{p}(\vec{V}) = \tilde{p}(\vec{e}_j V^j) = V^j \tilde{p}(\vec{e}_j) = V^j p_j, \quad p_j = \tilde{p}(\vec{e}_j). \quad (4.28)$$

This defines the components p_j of the form. Another notation often used is equivalent to (4.28), but emphasizes the symmetry between the vector space and the dual 1-form space,

$$\tilde{p}(\vec{V}) = \langle \tilde{p}, \vec{V} \rangle = V^j p_j. \quad (4.29)$$

This idea is of course familiar in elementary matrix theory, where row vectors form a dual space to column vectors, and map them into single numbers; the column vectors could also be thought of as mapping row vectors into scalars.

Being in a linear vector space the 1-forms will have a basis, which we denote $\tilde{\omega}^m$. We assume the expansion coefficients are the components defined in (4.28), so $\tilde{p} = p_m \tilde{\omega}^m$. Then we see that the basis 1-forms and basis vectors must obey an

orthogonality relation inferred by (4.28),

$$\tilde{p}(\vec{V}) = (p_m \tilde{\omega}^m)(\vec{e}_j V^j) = p_m V^j \tilde{\omega}^m(\vec{e}_j) = p_j V^j \text{ so } \tilde{\omega}^m(\vec{e}_j) = \delta_j^m. \quad (4.30)$$

That is, we have set up the basis forms to obey the orthogonality relation in (4.30).

One special 1-form is of particular interest and leads to a curious notation. In the context of forms the “gradient” of a function ϕ is defined as a form having components which are the usual partial derivatives $\phi_{,k}$; the gradient is thus

$$\tilde{d}\phi \equiv \phi_{,k} \tilde{\omega}^k. \quad (4.31)$$

The gradient 1-form of the coordinate x^i is then given by

$$\tilde{d}x^i = x^i_{,k} \tilde{\omega}^k = \frac{\partial x^i}{\partial x^k} \tilde{\omega}^k = \delta_k^i \tilde{\omega}^k = \tilde{\omega}^i, \quad \tilde{d}x^i = \tilde{\omega}^i, \quad (4.32)$$

Thus the basis may be expressed as the set of coordinate gradients. This implies that we may rewrite (4.31) as

$$\tilde{d}\phi = \phi_{,k} \tilde{d}x^k \quad (4.33)$$

which looks like the analogous elementary calculus expression, but is a relation between 1-forms. It is clear from (4.27) and (4.33) that the 1-forms should transform covariantly.

Example 4.3 Let us return to the polar coordinate system in Example 4.1 and obtain the relevant vectors and forms. From Fig. 4.2 it is clear that the coordinate basis vectors in the radial and angle directions must be

$$\vec{e}_\rho = a(\cos \varphi \vec{e}_x + \sin \varphi \vec{e}_y), \quad \vec{e}_\theta = b(-\sin \varphi \vec{e}_x + \cos \varphi \vec{e}_y). \quad (4.34)$$

where a and b are some constants. The basis vectors can have any length we choose.

According to (4.19) the metric is then,

$$g_{\mu\nu} = \begin{pmatrix} a^2 & 0 \\ 0 & b^2 \end{pmatrix}. \quad (4.35)$$

If we wish this to be the metric for flat Euclidean 2-space as in (4.18) we choose $a = 1$ and $b = \rho$ and have

$$\vec{e}_\rho = \cos \varphi \vec{e}_x + \sin \varphi \vec{e}_y, \quad \vec{e}_\theta = \rho(-\sin \varphi \vec{e}_x + \cos \varphi \vec{e}_y). \quad (4.36)$$

To obtain the basis 1-forms we may use the transformation rule (4.27) since the upper index position determines the transformation; alternatively we may use the demand that the forms obey the orthogonality relation (4.30). The result is

$$\tilde{dr} = \cos \varphi \tilde{dx} + \sin \varphi \tilde{dy}, \quad \tilde{d\varphi} = -\frac{1}{\rho}(\sin \varphi \tilde{dx} + \cos \varphi \tilde{dy}). \quad (4.37)$$

We will later further discuss the normalization of the 1-forms.

4.4 Tensors, Component View

We continue in this section with the classic component view of vectors and tensors as indexed arrays. This section consists largely of a set of theorems which are proved by a relatively simple algebraic process often called *index juggling*. It should become clear that after some practice the balancing of the indices does much of the work for us.

To define a tensor we generalize the idea of a vector as defined as an n -tuple with a well-defined transformation between coordinate systems: a tensor is defined as a set of quantities with any number of indices, which transforms according to

$$\bar{T}^{l\dots}_{m\dots} = \frac{\partial \bar{x}^l}{\partial x^q} \dots \frac{\partial x^n}{\partial \bar{x}^m} \dots T^{q\dots}_{n\dots}, \quad \text{tensor components.} \quad (4.38)$$

The total number of indices is referred to as the *rank*; some of the indices may be upper, or contravariant, and others may be lower, or covariant. The number of such indices is written as (M, N) . Thus for example a vector is a first rank tensor and $(1, 0)$. Another example is $V^j W_q$, which is a second rank tensor and $(1, 1)$.

From this tensor definition many simple but powerful theorems follow. We have already introduced and proved two of them in Sect. 4.2: Theorem 1 concerned the Jacobian matrices and Theorem 2 the invariance of the inner product of vectors. Let us continue to more such theorems.

Theorem 3 *To contract a tensor we set an upper index equal to a lower index and sum, which gives another tensor; for example one contraction of $T^{\alpha\beta}_{\lambda\gamma}$ is $T^{\alpha\beta}_{\beta\gamma} = S^{\alpha}_{\gamma}$. Contraction of a rank r tensor produces a rank $r - 2$ tensor. Consider the above 4th rank tensor as an example. Then the contracted object transforms as*

$$\begin{aligned} \bar{T}^{\alpha\beta}_{\beta\gamma} &= \frac{\partial \bar{x}^\alpha}{\partial x^\omega} \frac{\partial \bar{x}^\beta}{\partial x^\sigma} \frac{\partial x^\lambda}{\partial \bar{x}^\beta} \frac{\partial x^\eta}{\partial \bar{x}^\gamma} T^{\omega\sigma}_{\lambda\eta} = \frac{\partial \bar{x}^\alpha}{\partial x^\omega} \delta^\lambda_\sigma \frac{\partial x^\eta}{\partial \bar{x}^\gamma} T^{\omega\sigma}_{\lambda\eta} \\ &= \frac{\partial \bar{x}^\alpha}{\partial x^\omega} \frac{\partial x^\eta}{\partial \bar{x}^\gamma} T^{\omega\sigma}_{\sigma\eta}. \end{aligned} \quad (4.39)$$

This is the transformation law of a second rank tensor. It is clear from this how the general case works. Note that there may be several different contractions of a tensor, such as $T^{\omega\sigma}{}_{\sigma\eta}$ and $T^{\sigma\omega}{}_{\sigma\eta}$.

Theorem 4 *The direct product of tensors is a tensor of higher rank; for example $V^\mu W^\tau = T^{\mu\tau}$ is a second rank tensor. The proof of this is left as an easy exercise.*

Theorem 5 *If the metric transforms as a covariant tensor of rank 2 then the line element is invariant; this is what we originally postulated the line element should be. The theorem follows from Theorems 3 and 4 above, but it is so important that we work it out explicitly. The metric is assumed to be a second rank covariant tensor, so*

$$\begin{aligned}\bar{g}_{\alpha\beta} &= \frac{\partial x^\lambda}{\partial \bar{x}^\alpha} \frac{\partial x^\eta}{\partial \bar{x}^\beta} g_{\lambda\eta}, \text{ so } d\bar{s}^2 = \bar{g}_{\alpha\beta} d\bar{x}^\alpha d\bar{x}^\beta = \frac{\partial x^\lambda}{\partial \bar{x}^\alpha} \frac{\partial x^\eta}{\partial \bar{x}^\beta} \frac{\partial \bar{x}^\alpha}{\partial x^\omega} \frac{\partial \bar{x}^\beta}{\partial x^\sigma} g_{\lambda\eta} dx^\omega dx^\sigma \\ &= \delta_\omega^\lambda \delta_\sigma^\eta g_{\lambda\eta} dx^\omega dx^\sigma = g_{\lambda\eta} dx^\lambda dx^\eta.\end{aligned}\quad (4.40)$$

It is important to emphasize that the metric will not generally have the same form in the barred system as in the unbarred system. This is in distinction to special relativity where we carefully limited ourselves to transformations for which the metric did not change—the Lorentz group.

Theorem 6 *If the metric has an antisymmetric part it does not contribute to the line element. We have already mentioned this following the introduction of the line element (4.4) in Sect. 4.1, and also in Exercise 4.3. Because of this we always assume that the metric is symmetric.*

Theorem 7 *The symmetry character of a tensor is an invariant property. We illustrate this by showing that if a 2nd rank tensor is symmetric in one system it must be symmetric in another.*

$$\bar{T}^{\alpha\beta} = \frac{\partial \bar{x}^\alpha}{\partial x^\omega} \frac{\partial \bar{x}^\beta}{\partial x^\sigma} T^{\omega\sigma} = \frac{\partial \bar{x}^\alpha}{\partial x^\omega} \frac{\partial \bar{x}^\beta}{\partial x^\sigma} T^{\sigma\omega} = \bar{T}^{\beta\alpha}.\quad (4.41)$$

The general case is clear from this.

Theorem 8 *If a tensor equation is true in one system of coordinates then it is true in all systems. We illustrate this with the following equation involving a scalar, a tensor, and two vectors,*

$$T^{\mu\nu} = \phi V^\mu U^\nu.\quad (4.42)$$

Assume this is true in the unbarred system. In the barred system the two transformed tensors are

$$\bar{T}^{\alpha\beta} = \frac{\partial \bar{x}^\alpha}{\partial x^\mu} \frac{\partial \bar{x}^\beta}{\partial x^\nu} T^{\mu\nu} \quad \text{and} \quad \phi \bar{V}^\alpha \bar{U}^\beta = \frac{\partial \bar{x}^\alpha}{\partial x^\mu} \frac{\partial \bar{x}^\beta}{\partial x^\nu} \phi V^\mu U^\nu.\quad (4.43)$$

Because of (4.42) these are equal. The general case of any tensor equation is clear from this example. This theorem is the basis of a very powerful method of proof for tensor equations: they need only be proved in one convenient coordinate system, and are then automatically true in any coordinate system. An equation of this type between tensors is called *form invariant*, or *covariant* since the two sides vary together under the transformation.

Theorem 9 *Any tensor can be expanded as the sum of outer products of vectors.* As usual we illustrate this with a special case, a second rank (2, 0) tensor $T^{\mu\nu}$. To construct the expansion in an n -dimensional space choose one coordinate system, and in that system set up n contravariant vectors defined by $U_j^\alpha = \delta_j^\alpha$, where j labels which vector and α labels the components. Next, in that coordinate system define a set of scalars equal to the components of the tensor, that is $a_{ij} = T^{ij}$. Then obviously, by construction

$$T^{\alpha\beta} = \sum_{i,j} a_{ij} \delta_i^\alpha \delta_j^\beta = \sum_{i,j} a_{ij} U_i^\alpha U_j^\beta. \quad (4.44)$$

We have thereby constructed the desired expansion in the chosen coordinate system. Most important, we have defined the U_j^α to be vectors and the a_{ij} to be scalars, with values given in the chosen system and defined in another system by the transformation laws. Thus (4.42) is a tensor equation and true in any coordinate system, so the expansion is generally valid. As might be expected this theorem is often useful in proving other theorems. We have proven it using n vectors in n -dimensional space; this is not the *minimum* number of vectors in the tensor expansion, but suffices to prove most theorems of interest.

Theorem 10 (The Quotient Theorem) *Suppose we have an array $T^{\alpha\beta}$ and we are given that for any vector V_β the array*

$$S^\alpha = T^{\alpha\beta} V_\beta, \quad (4.45)$$

is a vector; then the array $T^{\alpha\beta}$ must be a tensor.

To prove this theorem express S^α and V_β in terms of vectors in the barred system, and write the above as

$$\frac{\partial x^\alpha}{\partial \bar{x}^\lambda} \bar{S}^\lambda = T^{\alpha\beta} \left(\frac{\partial \bar{x}^\tau}{\partial x^\beta} \bar{V}_\tau \right). \quad (4.46)$$

Now multiply and contract both sides of this with $\partial \bar{x}^\omega / \partial x^\alpha$ and use Theorem 1 to obtain

$$\frac{\partial \bar{x}^\omega}{\partial x^\alpha} \frac{\partial x^\alpha}{\partial \bar{x}^\lambda} \bar{S}^\lambda = \bar{S}^\omega = \left(T^{\alpha\beta} \frac{\partial \bar{x}^\omega}{\partial x^\alpha} \right) \left(\frac{\partial \bar{x}^\tau}{\partial x^\beta} \bar{V}_\tau \right) = \bar{T}^{\omega\tau} \bar{V}_\tau, \quad (4.47)$$

which is the equation in the barred frame. Now since the vector $\bar{V}_\tau$ is arbitrary its coefficients in the last expression must be equal, so we have

$$\frac{\partial \bar{x}^\omega}{\partial x^\alpha} \frac{\partial \bar{x}^\tau}{\partial x^\beta} T^{\alpha\beta} = \bar{T}^{\omega\tau}. \quad (4.48)$$

This is the transformation law of a tensor so the theorem is proven for this special case. The general case is when the array is any rank and the vector is arbitrary, and the proof goes through as above.

Theorem 11 *Our next theorem is very simple but often useful. If a tensor is zero in one coordinate system then it must be zero in all coordinate systems.* This is obvious from the definition (4.38).

The existence of a metric tensor allows us to associate a *covariant* vector with any *contravariant* vector. As we discussed in Sect. 4.3 we lower an index according to

$$V_\alpha = g_{\alpha\beta} V^\beta. \quad (4.49)$$

By the above Theorems 3 and 4 this is indeed a covariant vector. As discussed in Sect. 4.3 we define the inverse metric array $g^{\lambda\tau}$ by

$$g^{\lambda\tau} g_{\tau\nu} = \delta_\nu^\lambda, \quad (4.50)$$

and use it to raise an index as follows

$$g^{\mu\rho} V_\rho = V^\mu. \quad (4.51)$$

A simple but important theorem tells us that the raising and lowering operations are consistent inverses of each other.

Theorem 12 *The Kronecker delta transforms as a (1, 1) 2nd rank tensor; it is thus peculiar in that its components are the same in all systems.* Proof is straightforward with the use of Theorem 1.

$$\bar{\delta}_\beta^\alpha = \frac{\partial \bar{x}^\alpha}{\partial x^\rho} \frac{\partial x^\omega}{\partial \bar{x}^\beta} \delta_\omega^\rho = \frac{\partial \bar{x}^\alpha}{\partial x^\rho} \frac{\partial x^\rho}{\partial \bar{x}^\beta} = \delta_\beta^\alpha. \quad (4.52)$$

Because of this the relation defining the inverse metric assures us that it is a second rank contravariant tensor according to the Quotient Theorem. As is clear from above, the operations of raising and lowering are consistent: if we first lower and then raise an index we regain the original vector.

It is evident that we may raise or lower an index in any tensor in exactly the same way as with vectors. For example we may form

$$T^{\alpha\beta} g_{\beta\kappa} = T^\alpha_\kappa. \quad (4.53)$$

When we lower or raise an index we usually retain the same name for the tensor; hence we may think of a tensor as being expressible in terms of its *contravariant components or covariant components* or any combination of the 2. This is in accord with the abstract view we will address in the next section.

From the defining relation of the inverse metric in (4.50) the Kronecker delta is the mixed metric tensor so we may express it as $g^\alpha_\beta = \delta^\alpha_\beta$; it is an unusual tensor in that it is the same in all coordinate systems as we noted above.

4.5 Tensors, Abstract View

As with vectors we may view tensors as abstract objects instead of from the classic component point of view discussed in the previous section. In this abstract approach an (M, N) tensor is defined to linearly map M vectors and N 1-forms to the reals. For example a $(0, 2)$ tensor T operates as a linear map on vectors $\vec{V}, \vec{W}$ as follows

$$T(\vec{V}, \vec{W}) = T(V^\beta \vec{e}_\beta, W^\mu \vec{e}_\mu) = V^\beta W^\mu T(\vec{e}_\beta, \vec{e}_\mu) \equiv V^\beta W^\mu T_{\beta\mu} \quad (4.54)$$

The components $T_{\beta\mu}$ defined here are the same as we discussed in the previous section. Thus a vector is also a $(1, 0)$ tensor and a 1-form is also a $(0, 1)$ tensor. The metric is the most important special case of a $(0, 2)$ tensor, so we explicitly note its operation in terms of components

$$g(\vec{V}, \vec{W}) = V^\beta W^\mu g_{\beta\mu} \quad (4.55)$$

Let's look at another example of a $(0, 2)$ tensor. Define the direct product of two 1-forms as something that operates linearly on two vectors to give a real in the following natural way,

$$\tilde{p} \otimes \tilde{q} = \text{direct product of 1-forms, } \tilde{p} \otimes \tilde{q}(\vec{V}, \vec{W}) = \tilde{p}(\vec{V})\tilde{q}(\vec{W}) \quad (4.56)$$

That is, the first factor in the direct product operates on the first vector and the second factor in the direct product operates on the second vector. The direct product in (4.56) is thus a $(0, 2)$ tensor. It should be clear that we can extend the definition to the direct product of any number M of 1-form factors to produce a $(0, M)$ tensor and so forth.

Recall that we discussed in Sect. 4.3 a basis for 1-forms which we denoted as $\tilde{\omega}^\alpha$. We can similarly show that there exists a basis for the product of two 1-forms or $(0, 2)$ tensors. Indeed the basis is a linear combination of the direct product of the $\tilde{\omega}^\alpha$. We write that linear combination as

$$f = f_{\alpha\beta} \tilde{\omega}^\alpha \otimes \tilde{\omega}^\beta. \quad (4.57)$$

To see that this is indeed a basis we verify that it produces the same result (4.54) when operating on two vectors by writing out its operation

$$f(\vec{V}, \vec{W}) = f_{\alpha\beta} \tilde{\omega}^\alpha \otimes \tilde{\omega}^\beta (\vec{V}, \vec{W}) = f_{\alpha\beta} \tilde{\omega}^\alpha (\vec{V}) \tilde{\omega}^\beta (\vec{W}) = f_{\alpha\beta} V^\alpha W^\beta. \quad (4.58)$$

It should be clear that we can extend this idea to any number of factors in the direct product and thereby have a basis for $(0, M)$ tensors.

Recall that the coordinate basis may be written in terms of the gradients of the coordinates as in (4.31). This allows us to write a curious and useful expression for the metric tensor from (4.58),

$$g = g_{\alpha\beta} \tilde{dx}^\alpha \otimes \tilde{dx}^\beta. \quad (4.59)$$

This looks like the expression for the line element but is a relation between forms.

From the above definitions and (4.59) a $(0, 2)$ we see that a tensor, such as the metric, can also be viewed as producing a 1-form from a vector if we leave the second space blank, or $g(\vec{V}, -)$; this maps vectors to scalars according to (4.59) as follows

$$\begin{aligned} g(\vec{V}, -) &= g_{\alpha\beta} \tilde{dx}^\alpha \otimes \tilde{dx}^\beta (\vec{V}, -) = g_{\alpha\beta} \tilde{dx}^\alpha (\vec{V}) \tilde{dx}^\beta \\ &= g_{\alpha\beta} V^\alpha \tilde{dx}^\beta = V_\beta \tilde{dx}^\beta. \end{aligned} \quad (4.60)$$

That is, the metric lowers the index to produce the components of the 1-form.

Finally, it is now rather obvious how to define a tensor in general in terms of the coordinate basis vectors and basis forms: an (M, N) tensor is a linear mapping of N vectors and M 1-forms to the scalars; it may be expanded in terms of the bases and components as

$$\begin{aligned} T &= T^{\alpha\dots}_{\beta\dots} (\tilde{e}_\alpha \otimes \dots) (\tilde{dx}^\beta \otimes \dots), \\ \otimes &= \text{direct product}. \end{aligned} \quad (4.61)$$

The direct product of the basis vectors in the above is the obvious analog of the direct product of the basis 1-forms. The relation between the abstract tensor and its components in terms of the coordinate basis vectors and 1-forms is thus fundamental and clear.

4.6 Tetrads and n -Trads

In general relativity we often find it useful to use tetrads, a set of four basis vectors that forms an orthonormal basis as in special relativity. This sets up a reference frame at a point that is analogous to the reference frame of special relativity. The tetrad

differs from the set of coordinate basis vectors in that it is normalized and need not align with the coordinate axes. More generally, in n dimensions we define an n -trad, a set of n basis vectors $\vec{e}_a$ oriented and normalized so that

$$\vec{e}_a \cdot \vec{e}_b = \eta_{ab}, \quad (4.62)$$

where the η_{ab} matrix is chosen for convenience. It is usually taken to be the constant Lorentz metric in relativity theory but may be any constant matrix such as the Kronecker delta as needed in other situations; we refer to it as the n -trad metric. In this section the n -trads will be labeled with lower Latin indices early in the alphabet like b , and the space indices will usually be Greek.

Notice that the local Lorentz frame we previously discussed is essentially the same as the frame provided by the tetrads. Indeed it is possible to develop the theory of tetrads based on the transformation to the local Lorentz frame, although we will not do that here (Lawrie 1990).

In this section we will denote the coordinate basis as $\vec{g}_\beta$ to distinguish it from the n -trad basis $\vec{e}_a$, and it will be labeled with Greek indices. The n -trad may be expanded in terms of the coordinate basis as

$$\vec{e}_a = e_a^\beta \vec{g}_\beta, \quad e_a^\beta = n\text{-trad components in coordinate basis.} \quad (4.63)$$

This gives a beautiful relation for the n -trad metric in terms of the metric,

$$\begin{aligned} \eta_{ab} &= \vec{e}_a \cdot \vec{e}_b = (e_a^\beta \vec{g}_\beta) \cdot (e_b^\mu \vec{g}_\mu) = e_a^\beta e_b^\mu (\vec{g}_\beta \cdot \vec{g}_\mu) = e_a^\beta e_b^\mu g_{\beta\mu}, \\ \eta_{ab} &= e_a^\beta e_b^\mu g_{\beta\mu} \end{aligned} \quad (4.64)$$

The last expression may also be inverted to give the metric in terms of the n -trad metric, a very useful result. To do this we solve (4.64) for the metric $g_{\beta\mu}$; define the inverse of the n -trad component matrix e_b^μ and label it with a bow, according to

$$\check{e}_\gamma^a = \text{inverse of } e_a^\beta, \quad \text{so} \quad \check{e}_\gamma^a e_a^\beta = \delta_\gamma^\beta. \quad (4.65)$$

Using this we can solve (4.64) for the metric directly as follows

$$\begin{aligned} \check{e}_\gamma^a \eta_{ab} \check{e}_\sigma^b &= \check{e}_\gamma^a (e_a^\beta g_{\beta\mu} e_b^\mu) \check{e}_\sigma^b = \delta_\gamma^\beta g_{\beta\mu} \delta_\sigma^\mu = g_{\gamma\sigma}, \\ g_{\gamma\sigma} &= \check{e}_\gamma^a \eta_{ab} \check{e}_\sigma^b \end{aligned} \quad (4.66)$$

Thus in relativity theory, with the tetrad matrix equal to the Lorentz metric, the inverse $\check{e}_\gamma^a$ of the component matrix serves, loosely speaking, as a sort of matrix “square root” of the metric.

Sometimes the overhead bow on $\check{e}_\gamma^a$ in (4.66) is omitted and the index position reminds us that it is the *inverse* of the n -trad component matrix, that is with Latin index up and Greek index down. This is analogous to the use of the same symbol for the metric and its inverse, with the index position indicating which is which.

Both the coordinate basis $\vec{g}_\beta$ and the n -trad $\vec{e}_b$ can serve as bases in which to expand a given vector. Thus we may write

$$\vec{V} = V^\beta \vec{g}_\beta = V^c \vec{e}_c, \quad (4.67)$$

where as before we have denoted the vector components by Greek indices and the n -trad components by Latin indices. It is useful to relate the two types of components. We can express the n -trads in terms of the coordinate basis using (4.63) and obtain

$$V^\beta \vec{g}_\beta = V^c e_c^\beta \vec{g}_\beta \quad \text{so} \quad V^\beta = V^c e_c^\beta. \quad (4.68)$$

Thus the components in the two systems are simply related by the tetrad component matrix. The last relation above is easily inverted to give

$$V^b = V^\gamma \check{e}_\gamma^b \quad (4.69)$$

There are many simple algebraic relations like this that can be obtained by straightforward algebra. For example it is easy to see that n -trad indices are raised and lowered with the tetrad matrix, squares of vectors are n -trad squares and so forth.

Example 4.4 To illustrate the ideas of coordinate bases and n -trads an example is in order. A simple and useful spherical metric for this is the following,

$$ds^2 = F(r)dr^2 + r^2(d\theta^2 + \sin^2\theta d\varphi^2), \quad F = \text{smooth function}. \quad (4.70)$$

The coordinate basis vectors then lie along the coordinate directions, are orthogonal, and are normalized with the metric according to (4.21),

$$\vec{g}_1 \cdot \vec{g}_1 = g_{11} = F, \quad \vec{g}_2 \cdot \vec{g}_2 = g_{22} = r^2, \quad \vec{g}_3 \cdot \vec{g}_3 = g_{33} = r^2 \sin^2\theta. \quad (4.71)$$

For maximum simplicity let us also put the 3-trad or triad vectors along the coordinate directions and of course normalize with the Kronecker delta rather than the Lorentz metric. That is the triad lies in the same direction as the coordinate basis but the normalization is different. The triad components thus must have the general form

$$e_1^\mu = (A, 0, 0), \quad e_2^\mu = (0, B, 0), \quad e_3^\mu = (0, 0, C), \quad (4.72)$$

and we need only determine the quantities A and B and C . For that we normalize the 3-trad using (4.63) and (4.64); for the first triad vector the normalization demand is

$$\vec{e}_1 \cdot \vec{e}_1 = (e_1^\beta \vec{g}_\beta) \cdot (e_1^\nu \vec{g}_\nu) = e_1^\beta e_1^\nu g_{\beta\nu} = A^2 F = 1. \quad (4.73)$$

Thus we have $A = 1/\sqrt{F}$. The B and C are determined in the same way and give e_c^μ and its inverse $\check{e}_\gamma^b$ as (4.74)

$$e_c^\mu = \begin{pmatrix} 1/\sqrt{F} & 0 & 0 \\ 0 & 1/r & 0 \\ 0 & 0 & 1/r \sin \theta \end{pmatrix}, \quad \check{e}_\sigma^b = \begin{pmatrix} \sqrt{F} & 0 & 0 \\ 0 & r & 0 \\ 0 & 0 & r \sin \theta \end{pmatrix}. \quad (4.74)$$

It is easy to see from this that the relation (4.66) giving the metric in terms of the triad vector array is satisfied: the metric is the square of $\check{e}_\sigma^b$.

Many of the metric tensors encountered in relativity are diagonal, but certainly not all of them. See Example 4.6 for a simple example of coordinate basis vectors and 2-trad or dyad relations.

Tetrads are often useful in general relativity since they provide a beautiful connection with special relativity, analogous to the transformation to the local Lorentz frame. For example they allow us to incorporate spin one half particles, described by spinors, into the general theory. The Dirac equation describing such spinors is intimately connected with representations of the Lorentz group so the tetrad formalism is natural for their study (Lawrie 1990).

4.7 Volume Elements

In general relativity we often need the integral of a scalar function, which itself is a scalar. The integral of a vector or tensor will not in general have a well-defined transformation law since it is not a quantity defined at a single point. Our task in this section is to obtain an expression for a volume element to be used when integrating a scalar function over all or part of a space.

The appropriate expression for a volume element in a general Riemann space can be obtained by first considering the special case of a diagonal metric and then generalizing to any metric using invariance arguments. For a diagonal metric in any number of dimensions we may write the line element as

$$ds^2 = \sum_i g_{ii} (dx^i)^2, \quad d\ell_i \equiv \sqrt{g_{ii}} dx^i = \text{physical distance in } i \text{ direction}. \quad (4.75)$$

That is, as we discussed in Sect. 4.3, the $d\ell_i$ is a physical distance interval. (We assume for the moment that g_{ii} is positive.) What is particularly nice is that this allows us to define a physically meaningful n -volume element in a clear and obvious

way, as the product of the physical distances,

$$dV_n \equiv d\ell \dots d\ell_n = \sqrt{g_{11} \dots g_{nn}} dx^1 \dots dx^n = \sqrt{|g|} d^n x, \dots \quad (4.76)$$

where $|g|$ denotes the determinant of the metric tensor. This last expression turns out to be general, except that one must use the absolute value of the determinant if the signature is negative.

To show that the expression (4.76) is the correct volume element we prove the following theorem.

Theorem *The object defined in (4.76) is an invariant in the sense that the integral of a scalar f over a given region is an invariant. We will show this in two dimensions, with the extension to any number of dimensions evident. The theorem in two dimensions says*

$$\int f(\bar{x}^1, \bar{x}^2) \sqrt{|\bar{g}|} d\bar{x}^1 d\bar{x}^2 = \int f(x^1, x^2) \sqrt{|g|} dx^1 dx^2. \quad (4.77)$$

The proof is in two parts. To first get the transformation of the metric determinant we write out the transformation of the metric and take the determinant of both sides to obtain

$$\begin{aligned} \bar{g}_{ij} &= \frac{\partial x^m}{\partial \bar{x}^i} \frac{\partial x^n}{\partial \bar{x}^j} g_{mn}, & |\bar{g}| &= \left| \frac{\partial x^m}{\partial \bar{x}^i} \right|^2 |g|, \\ \sqrt{|\bar{g}|} &= \left| \frac{\partial x^m}{\partial \bar{x}^i} \right| \sqrt{|g|} \equiv \left| \frac{\partial x}{\partial \bar{x}} \right| \sqrt{|g|}. \end{aligned} \quad (4.78)$$

We have again assumed that the metric determinant is positive. The vertical bars denote the determinant of the inverse Jacobian matrix, that is the inverse of the Jacobian. The indices have been dropped in the last expression since they are not needed, which is a common notation. Anything that transforms like $\sqrt{|\bar{g}|}$ in (4.78) is called a *scalar density*.

Next recall from integral calculus that the transformation of a surface area element involves the Jacobian and may be written as

$$d\bar{x}^1 d\bar{x}^2 = \left| \frac{\partial \bar{x}}{\partial x} \right| dx^1 dx^2. \quad (4.79)$$

Since the root metric determinant $\sqrt{|g|}$ transforms via the inverse Jacobian and the area element transforms via the Jacobian the product is an invariant

$$\sqrt{|\bar{g}|} d\bar{x}^1 d\bar{x}^2 = \sqrt{|g|} dx^1 dx^2, \quad (4.80)$$

and the theorem is proved. For the evident generalization to n dimensions we have

$$dV_n = \sqrt{|g|} dx^1 \dots dx^n \text{ invariant } n\text{-volume element.} \quad (4.81)$$

The only caveat needed is that for the general case the metric determinant in (4.81) can be negative, as it generally is in relativity, and we must then use the absolute value for $|g|$. Note also that the scalar in (4.81) reduces to the obviously correct expression in the local frame where the metric is the Cayley-Sylvester canonical form with diagonal elements equal to 1 or -1 .

Example 4.5 Volume elements are often fairly easy to calculate. Here are some examples with diagonal metrics. For the curved 2-space with polar coordinates in Example 4.2 the volume element is

$$dV_2 = \sqrt{1 + f'^2} \rho d\rho d\varphi, \quad \text{polar.} \quad (4.82)$$

For cylindrical coordinates in flat 3-space the volume element is

$$dV_3 = \rho d\rho d\varphi dz \quad \text{cylindrical.} \quad (4.83)$$

For spherical coordinates in 3-space, the volume element in Example 4.4 is

$$dV_3 = \sqrt{F} r^2 \sin \theta dr d\theta d\varphi \quad \text{spherical.} \quad (4.84)$$

For Minkowski spacetime in Cartesian or spherical coordinates,

$$dV_4 = c dt dx dy dz = r^2 \sin^2 \theta c dt dr d\theta d\varphi \quad \text{flat spacetime.} \quad (4.85)$$

Non-diagonal metrics are also straight forward to analyze but somewhat more subtle, as we show in the following example.

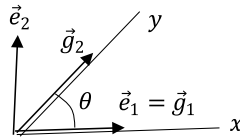


Fig. 4.6 Coordinate basis and dyad in a tilted coordinate x, y system. All the vectors are normalized to unit length

Example 4.6 As the simplest example of a non-diagonal metric consider 2-dimensional Euclidean space with Cartesian-like coordinates, but with a tilted y axis as in Fig. 4.6.

From the figure the line element and the metric tensor are

$$ds^2 = dx^2 + dy^2 + 2 \cos \theta dx dy, \quad g_{ik} = \vec{g}_i \cdot \vec{g}_k = \begin{pmatrix} 1 & \cos \theta \\ \cos \theta & 1 \end{pmatrix}. \quad (4.86)$$

To get the metric in another way we express the coordinate basis in terms of the orthonormal dyad,

$$\vec{g}_1 = \vec{e}_1, \quad \vec{g}_2 = \cos \theta \vec{e}_1 + \sin \theta \vec{e}_2 \quad \text{with } \delta_{jk} = \vec{e}_j \cdot \vec{e}_k \quad (4.87)$$

and find the same metric tensor (4.86).

Lastly, we work out the 2-volume element. The metric determinant from (4.86) and the 2-volume or area are

$$|g| = \sin^2 \theta, \quad dV_2 = \sin \theta dx dy. \quad (4.88)$$

This agrees with simple geometry and Fig. 4.6.

In this section we have chosen to develop the theory of invariant volume elements and integrals in a simple way, depending on invariance arguments. However if one pursues these ideas further he is lead naturally into the theory of p -forms (Ohanian 1994; Misner 1973). When antisymmetric 2-tensors are considered such forms become very useful. In this book we will develop the physics of gravity and cosmology without the use of such p -forms but will discuss them very briefly in Appendix 2 in Chap. 6.

Appendix 1: Differential Manifolds

The term manifold occurs in more mathematically oriented work. A manifold is an open collection of points P with useful properties for physics applications; because the collection is open there will be by definition a region around each point that is also in the manifold. The points in a 1-dimensional manifold are in one to one correspondence with an open set of the reals; an open set of the reals is defined as a union of open intervals. Similarly the points in a 2-dimensional manifold are in one to one correspondence with a pair of reals, and so forth for any dimension n . Thus an n -dimensional manifold is an open set of points that can be labeled by an n -tuple of reals in an open region, that is by coordinates.

A function of the manifold points P can be defined in terms of a function of the coordinates as $f(P) = f(x^k)$. Continuous and differentiable functions are naturally of particular usefulness.

In general the labeling of the points in a manifold is not unique and several coordinate systems may be used to label a region of the manifold; they are often denoted as unprimed and primed, or unbarred and barred as we have done. If there is a differentiable and invertible transformation $\bar{x}^k = \bar{x}^k(x^i)$ between any two such coordinate systems we say that the manifold is differentiable. This means that a differentiable function of the points in the manifold corresponds to a continuous function in both coordinate systems.

Thus, in short, a manifold is simply the kind of space physics has used for centuries, defined a bit more carefully with an emphasis on coordinate systems and open sets.

Appendix 2: The Signature Theorem in Two Dimensions

The Signature Theorem states that one can find a linear transformation to a coordinate system in which the metric tensor is diagonal and has 1, -1 , or 0 as diagonal elements. We will illustrate the proof in two dimensions with signature (1, 1) for simplicity. This is clearly a matrix problem so we will use matrix notation rather than index notation. We need deal only with a single point. The transformation between the original system and a new barred system we write in matrix form as

$$x = D\bar{x}, \quad (4.89)$$

where D is a matrix to be determined. The metric G transforms as a second rank tensor so its transformation in matrix form is

$$\bar{G} = D^T G D, \quad G = \begin{pmatrix} g_{11} & g_{12} \\ g_{12} & g_{22} \end{pmatrix}. \quad (4.90)$$

We first make the metric diagonal by choosing the transformation matrix D with a single parameter b to be determined,

$$D = \begin{pmatrix} 1 & 0 \\ b & 1 \end{pmatrix}. \quad (4.91)$$

After the transformation the metric becomes

$$\bar{G} = \begin{pmatrix} g_{11} + 2bg_{12} + b^2g_{22} & g_{12} + bg_{22} \\ g_{12} + bg_{22} & g_{22} \end{pmatrix}. \quad (4.92)$$

We make this diagonal by choosing $b = -g_{12}/g_{22}$. Then we have

$$\overline{G} = \begin{pmatrix} g_{11} - (g_{12})^2/g_{22} & 0 \\ 0 & g_{22} \end{pmatrix} = \begin{pmatrix} \bar{g}_{11} & 0 \\ 0 & \bar{g}_{22} \end{pmatrix}. \quad (4.93)$$

Notice that the 1, 1 element of the matrix $\overline{G}$ is the determinate of G divided by g_{22} ; we will assume this is positive so that the signature will be (1, 1). (The reader should work out the case where it is negative and the signature is (1, -1) as in Exercise 4.7)

We next apply a second linear transformation to stretch the coordinates and make both diagonal elements of the metric equal to 1. Specifically this is done with the obvious stretching matrix

$$\begin{pmatrix} \sqrt{\bar{g}_{11}} & 0 \\ 0 & \sqrt{\bar{g}_{22}} \end{pmatrix} \quad (4.94)$$

The metric is then the 2-dimensional unit matrix as desired. We have thus obtained the Cayley-Sylvester canonical form by two successive linear transformations.

We emphasize again that the manipulations apply at a single point P . At a different point the metric will in general not have the canonical form. For n dimensions the theorem is still relatively easy to prove (Courant 1937; Perlis 1952).

Exercises

- 4.1 Suppose that S_{ij} is a matrix array that is symmetric in its indices, and that A^{ij} is an antisymmetric array. Show that the product $S_{ij}A^{ij}$ is zero.
- 4.2 Show that one may express any second rank matrix as the sum of a symmetric and an antisymmetric matrix.
- 4.3 From the above two exercises show that if the metric is not symmetric then only the symmetric part of it matters in the line element, that is $(g_{ij} + g_{ji})/2$. This is one reason why we always assume the metric is symmetric.
- 4.4 Work out the metric (4.8) in Example 4.1 for plane polar coordinates using the transformation law (4.40) from Cartesian coordinates to polar coordinates and see that you get the same result.
- 4.5 Work out the metric for spherical coordinates (r, θ, φ) in Euclidean 3-space. First do this by using a picture of a small box analogous to that in Fig. 4.3. Then do it by transforming the metric from Cartesian coordinates (3 by 3 identity), using the transformation law (4.40). Which is easier?
- 4.6 Work out the metric on the curved 2-surface of a sphere of radius R for a number of coordinate systems. First, use Cartesian coordinates with the constraint $R^2 = x^2 + y^2 + z^2$ and express the line element in terms of x and y . Secondly, do it with cylindrical coordinates following our discussion in Example 4.2. Finally, do it with spherical coordinates. You should notice how a coordinate system with the appropriate symmetry makes the process simple.
- 4.7 Go through Sect. 4.7 for the case in which the metric determinant is negative. Similarly go through the proof of the Signature Theorem in two dimensions for the case where the signature is (1, -1) so the metric determinant is negative. What difficulties occur if the signature has a 0?

- 4.8 Repeat the flat space analysis in Example 4.3 but do it for the curved space in Example 4.2 with the metric (4.11). Work out the coordinate basis vectors and 1-form basis. Write out the matrices e_c^μ and $\check{e}_\sigma^b$.
- 4.9 In (4.29) we expressed in symmetric notation the operation of a 1-form in mapping vectors to the reals. Show (briefly) that we could equally well define vectors as mapping 1-forms to the reals, hence the symmetric notation.

Chapter 5

Affine Connections and Geodesics



Abstract In a general Riemann space the concepts of straight lines and parallel vectors must be generalized from those familiar in Euclidian geometry. The fundamental objects needed for the generalization are affine connections. With affine connections we are naturally led to a deeper view of spacetime and the behavior of objects in it.

5.1 Affine Connections, Component View

Most of our considerations in Chap. 4 involved vectors and tensors associated with a single point. Now we study how to compare vectors and tensors at different points in a Riemann space, and how to move them (Misner 1973; Adler 1975; Schutz 2009). This is necessary in order to study tensor fields, that is tensors defined as functions of position in regions of space; these fields may be denoted for example as

$$\phi(x^\mu) \text{ scalar, } V^\alpha(x^\mu) \text{ vector, } T^{\alpha\beta}(x^\mu) \text{ 2nd rank tensor.} \quad (5.1)$$

This is not a trivial process since vector spaces at different points in a Riemann space are a priori independent and any connection between them requires analysis. The key concept is that of affine connections, for which we will motivate a definition; then on the basis of the definition we may obtain their transformation law.

Much of the work in this chapter is based on the classic component view, but we will relate it to the abstract view in the last section.

Consider first a constant vector field in Euclidean 3-space with Cartesian coordinates; the definition of such a constant vector field is obviously that the components are constant, as shown in Fig. 5.1a. But it is also clear that in spherical coordinates constant components do *not* correspond to what we think of as a constant vector field, as in Fig. 5.1b. Clearly, we should not define a constant vector field as one with constant components. The terms “constant” and “parallel” remain to be defined precisely. The proper definition of a constant vector field will introduce the concept of affine connections as an elegant generalization of the notion of parallel vectors.

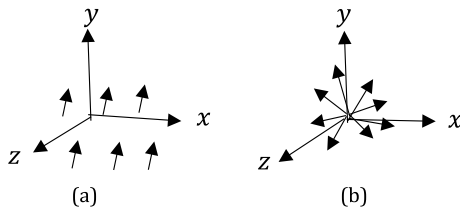


Fig. 5.1 **a** A vector field in Cartesian coordinates with constant components; the field is obviously constant and the vectors are parallel to each other. **b** A vector field in spherical coordinates with constant components, for example $(1, 0, 0)$; the field is not constant and the vectors are not parallel

We first motivate the definition with a special case: suppose that we are in a flat Euclidean space with Cartesian coordinates, but we wish to consider other coordinate systems as well, as in the example above. In the Cartesian system we take the definition of a constant field to be that the components are constant: they do not change as we go to a nearby point,

$$V^i(x^j) = \text{const.}, \quad dV^i = 0. \quad (5.2)$$

In another barred system that is not Cartesian the vector components and changes are easily obtained from the definition of a contravariant component vector in (4.14),

$$d\bar{V}^i = d\left(\frac{\partial \bar{x}^i}{\partial x^j} V^j\right) = \left(d\frac{\partial \bar{x}^i}{\partial x^j}\right) V^j = \left(\frac{\partial^2 \bar{x}^i}{\partial x^l \partial x^j} dx^l\right) V^j = \left(\frac{\partial^2 \bar{x}^i}{\partial x^l \partial x^j}\right) dx^l V^j \quad (5.3)$$

We wish to relate this change in the vector components to the coordinate differentials and the vector components expressed in the barred system by using the transformation equations (4.13) and (4.14); we find

$$\begin{aligned} V^j &= \frac{\partial x^j}{\partial \bar{x}^n} \bar{V}^n, \quad dx^l = \frac{\partial x^l}{\partial \bar{x}^k} d\bar{x}^k, \\ d\bar{V}^i &= \left(\frac{\partial^2 \bar{x}^i}{\partial x^j \partial x^l} \frac{\partial x^l}{\partial \bar{x}^k} \frac{\partial x^j}{\partial \bar{x}^n}\right) d\bar{x}^k \bar{V}^n \equiv -\bar{\Gamma}_{kn}^i d\bar{x}^k \bar{V}^n. \end{aligned} \quad (5.4)$$

Thus we see that in the non-Cartesian system the change in the vector components is of course not zero, but is a bilinear function of the coordinate differentials and the vector components; this linear relation leads to the name *coefficients of affine connection* given to the array $\bar{\Gamma}_{kn}^i$ defined in (5.4). They are often called affine connections or simply connections as we will usually do. Note from (5.4) that the connections in this example are symmetric in the lower two indices. The use of a minus sign in the definition is for later convenience when we define covariant derivatives.

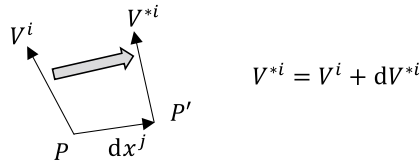


Fig. 5.2 Transplantation of a vector from a point to a nearby point. It allows the comparison of vectors at nearby points. The components of the vector will change according to (5.5)

The above example illustrates the idea of an affine connection but it is not general enough. It is only useful for the special spaces in which a global Cartesian coordinate system can be established; there are many spaces for which this is not the case, so we must treat the idea in more generality. Thus we *postulate* that in the space and coordinates considered there exists a set of affine connections which are functions of position, and a vector V^j is said to be *transplanted* by dx^i from a given point to a nearby point (see Fig. 5.2) if its components change according to

$$dV^{*i} = -\Gamma_{kn}^i dx^k V^n, \text{ law of vector transplantation.} \quad (5.5)$$

We must emphasize that (5.5) *defines* the vector transplanted from P to P' . If the vector is a field then its value at P' need not be the same as the transplanted vector at P' . Indeed this difference is central to the ideas of vector and tensor analysis in Chap. 6.

We motivated the transplantation law using a special case of a Euclidean space, but alternatively we could have simply postulated it ad hoc; it is clearly reasonable that the change in a vector should be proportional to the vector itself and to the distance over which it is transplanted. The transplantation law is very general and is central to the ideas of vector and tensor derivatives in Chap. 6.

The law of vector transplantation in (5.5) is presumed to hold in any of the Riemann spaces that we will consider. A space in which there are such connections is termed an affine space. From what we have said so far, the affine connections could be taken to have any values desired; alternatively they may be determined by some physical or geometric demand. In relativity theory we follow the latter course and in Sect. 5.3 will impose a geometric or physical demand to obtain the connections called the Christoffel connections.

In Sect. 5.6 on the abstract view we will look at the problem in another way and relate the affine connections to changes in the coordinate basis vectors.

5.2 Transformation of the Affine Connections

The law of vector transplantation introduced in (5.5) is extremely general since there are no restrictions on the connections. Remarkably, from only the above definition

of vector transplantation we may obtain the transformation law for the connections, which we will find are *not* tensors. Moreover several theorems that result from the transformation law are basic and important for both mathematics and physics.

To find the transformation law we make the natural demand that a vector remains a vector as it is transplanted to a nearby point: that is it must obey the transformation law (4.14) at both P and P' . The transplanted vector, at P' in the barred and unbarred coordinate systems, is

$$V^{*i} = V^i - \Gamma_{pq}^i dx^p V^q, \quad \bar{V}^{*j} = \bar{V}^j - \bar{\Gamma}_{mn}^j d\bar{x}^m \bar{V}^n. \quad (5.6)$$

Here the vector and connections on the right side of (5.6) are evaluated at P . The transformation matrix at P' may be gotten with a Taylor series expansion from that at P ,

$$\left(\frac{\partial \bar{x}^j}{\partial x^i} \right)_{P'} = \left(\frac{\partial \bar{x}^j}{\partial x^i} \right)_P + \frac{\partial}{\partial x^l} \left(\frac{\partial \bar{x}^j}{\partial x^i} \right)_P dx^l = \left(\frac{\partial \bar{x}^j}{\partial x^i} \right)_P + \left(\frac{\partial^2 \bar{x}^j}{\partial x^l \partial x^i} \right)_P dx^l. \quad (5.7)$$

We use these expressions and impose the vector transformation law on the vector at P' ,

$$\begin{aligned} \bar{V}^{*j} &= \left(\frac{\partial \bar{x}^j}{\partial x^l} \right)_{P'} V^{*l} \quad \text{so} \\ \bar{V}^j - \bar{\Gamma}_{mn}^j d\bar{x}^m \bar{V}^n &= \left[\left(\frac{\partial \bar{x}^j}{\partial x^i} \right)_P + \left(\frac{\partial^2 \bar{x}^j}{\partial x^l \partial x^i} \right)_P dx^l \right] [V^i - \Gamma_{pq}^i dx^p V^q] \\ &= \left(\frac{\partial \bar{x}^j}{\partial x^i} \right)_P V^i - \left(\frac{\partial \bar{x}^j}{\partial x^i} \right)_P \Gamma_{pq}^i dx^p V^q + \left(\frac{\partial^2 \bar{x}^j}{\partial x^l \partial x^i} \right)_P dx^l V^i. \end{aligned} \quad (5.8)$$

The first terms on each side of this equation cancel because V^i is a vector at P . We relabel the dummy indices and the remaining terms tell us that

$$-\bar{\Gamma}_{ml}^j d\bar{x}^m \bar{V}^l = \left[-\left(\frac{\partial \bar{x}^j}{\partial x^n} \right)_P \Gamma_{qi}^n + \left(\frac{\partial^2 \bar{x}^j}{\partial x^q \partial x^i} \right)_P \right] dx^q V^i. \quad (5.9)$$

Next, we express the vector and coordinate differential in the unbarred system in terms of those in the barred system using the vector transformation equations, and find

$$\begin{aligned} &-\bar{\Gamma}_{ml}^j d\bar{x}^m \bar{V}^l \\ &= \left[-\left(\frac{\partial \bar{x}^j}{\partial x^n} \frac{\partial x^q}{\partial \bar{x}^m} \frac{\partial x^i}{\partial \bar{x}^l} \right)_P \Gamma_{qi}^n + \left(\frac{\partial^2 \bar{x}^j}{\partial x^q \partial x^i} \right)_P \frac{\partial x^q}{\partial \bar{x}^m} \frac{\partial x^i}{\partial \bar{x}^l} \right] d\bar{x}^m \bar{V}^l. \end{aligned} \quad (5.10)$$

Finally we observe that we may impose this relation on any vector and any displacement; because of this the coefficients of the array $d\bar{x}^m \bar{V}^i$ on the two sides of (5.10) must be equal, so the affine connections must transform according to

$$\bar{\Gamma}_{ml}^j = \frac{\partial \bar{x}^j}{\partial x^n} \frac{\partial x^q}{\partial \bar{x}^m} \frac{\partial x^i}{\partial \bar{x}^l} \Gamma_{qi}^n - \frac{\partial^2 \bar{x}^j}{\partial x^q \partial x^i} \frac{\partial x^q}{\partial \bar{x}^m} \frac{\partial x^i}{\partial \bar{x}^l}. \quad (5.11)$$

We have dropped the subscript P which is no longer necessary since everything in the equation is evaluated at P . Notice that the first term in this relation is that of a tensor transformation as defined in (4.38), but the second term is inhomogeneous and independent of the connections. It is important that in general the affine connections do not transform as tensors.

Several interesting properties of the connections follow from the transformation law (5.11), which we will state as theorems.

Theorem 1 *Under the special case of linear transformations the affine connections do transform as tensors.* This follows since the second derivatives in (5.11) vanish for linear transformations.

Theorem 2 *If the affine connections are symmetric in their lower indices in one coordinate system then they are symmetric in all coordinate systems.* The proof is evident from the transformation law (5.11) since the second term is symmetric.

Theorem 3 *If the affine connections vanish in one coordinatesystem then they are symmetric in any coordinate system.* This is also evident from (5.11).

Theorem 4 (A beautiful and fundamental theorem of Weyl) *If the affine connections are symmetric then there exists a coordinatesystem in which they vanish.* We may prove this at the origin of the coordinate system without loss of generality. To prove the theorem consider a transformation of the form

$$\bar{x}^j = x^j + \frac{1}{2} A_{ik}^j x^i x^k. \quad (5.12)$$

Here A_{ik}^j is an array of constants to be determined. Then at the origin the following equations follow

$$\frac{\partial \bar{x}^j}{\partial x^i} = \delta_i^j, \quad \frac{\partial x^k}{\partial \bar{x}^n} = \delta_n^k, \quad \frac{\partial^2 \bar{x}^j}{\partial x^q \partial x^i} = \frac{1}{2} (A_{iq}^j + A_{qi}^j). \quad (5.13)$$

It then follows from the transformation law that at the origin the transformed connections are

$$\bar{\Gamma}_{ml}^j = \Gamma_{ml}^j - \frac{1}{2} (A_{ml}^j + A_{lm}^j). \quad (5.14)$$

Now we choose the array A_{ml}^j to be the negative of the affine connection array in the unbarred system, and thereby cause the connections at P to be zero in the barred system; we may do this *if and only if the connections are symmetric*. The coordinate system where the affine connections vanish is termed the *geodesic system* (Adler 1975).

Besides being elegant mathematics the Weyl Theorem has important implications for physics. We will see in Sect. 7.3 that in the geometric description of gravity the affine connections play a role analogous to Newtonian forces, and the Weyl Theorem thus tells us that gravitational effects may be transformed away at a point by a choice of coordinates! This is a very profound fact in general relativity theory and a cornerstone of the geometric view of gravity. As we will discuss in Sect. 7.2 equivalence principle (EP) experiments indicate that it is true in nature to an accuracy better than a part in 10^{13} (Wiki STEP). Because of this agreement with nature and because of mathematical elegance we will generally assume that the affine connections are symmetric.

If the affine connections are not taken to be symmetric a more general theory of gravity can be developed, the most well-known of which is the Einstein-Cartan theory. The effects of the non-symmetry of the connections are termed *torsion*. There is at present no experimental evidence for torsion to motivate such theories, but some theorists believe torsion may be necessary in a future theory of quantum gravity (Trautman 2006).

5.3 Parallel Displacement

The law of vector transplantation (5.5) introduced in the preceding section provides a way to compare vectors at different nearby points in space. By repeated iterations we could also compare vectors at widely separated points. Our considerations have so far been quite general and we made no assumptions about how the connections might be specified. We now specialize to obtain the specific connections used in relativity theory; this provides a strikingly elegant generalization of the idea of moving a vector parallel to itself in Euclidean geometry, and is called parallel displacement. Parallel displacement is basic to the idea and definition of space curvature that we will develop in Chap. 8. It also allows us to define geodesic curves, which are the generalization of straight lines to general Riemann spaces.

Suppose that we transplant two vectors to a nearby point using the law of vector transplantation. There is no a priori reason that the inner product of the two will remain unchanged; however we may consider this to be a naturally compelling *demand* to be imposed so as to make the transplantation analogous to the parallel displacement of vectors in Euclidean geometry. In the special case of Euclidean space the demand for parallelism implies that the lengths of various vectors and the angles between them remain unchanged as they are transplanted. We thus impose this demand and refer to this special case of vector transplantation as generalized parallel displacement, or simply *parallel displacement* for brevity. Remarkably, the connections are then

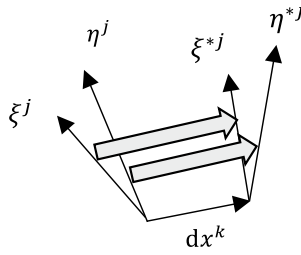


Fig. 5.3 In parallel displacement the two vectors are transplanted to a nearby point and we demand that the inner product of the two be unchanged

determined uniquely by the metric. Figure 5.3 shows the scenario for the parallel displacement of two vectors.

The derivation of the affine connections is conceptually simple and involves only slightly tedious algebra. The demand that the inner product of the two vectors be unchanged under vector transplantation may be expressed as

$$d(\xi^j \eta^k g_{jk})^* = 0, \quad (5.15)$$

where the change is that imposed by the vector transplantation law (5.5). This demand leads explicitly to

$$\begin{aligned} d(\xi^j \eta^k g_{jk})^* &= (d\xi^{*j})\eta^k g_{jk} + \xi^j (d\eta^{*k})g_{jk} + \xi^j \eta^k (dg_{jk}) \\ &= -(\Gamma_{pq}^j dx^p \xi^q)\eta^k g_{jk} - (\Gamma_{pq}^k dx^p \eta^q)\xi^j g_{jk} + (g_{jk,l} dx^l)\xi^j \eta^k \\ &= [g_{ik,l} - \Gamma_{li}^r g_{rk} - \Gamma_{lk}^r g_{ir}] dx^l \xi^i \eta^k = 0. \end{aligned} \quad (5.16)$$

(Notice the relabeling of dummy indices, or index juggling.) We emphasize that the change in the vectors is not due to any change in the value of vector *fields*, but is only the change associated with transplantation. The metric on the other hand changes because it is a tensor field. The last equation (5.16) is presumed to hold for any pair of vectors and any displacement, so the bracket on the last line must be zero, and we obtain the following relation between the metric and the affine connections,

$$g_{ik,l} - \Gamma_{li}^r g_{rk} - \Gamma_{lk}^r g_{ir} = 0. \quad (5.17a)$$

The last relation can be solved for the affine connections by index juggling. We first repeat it twice with the names of the indices permuted,

$$g_{kl,i} - \Gamma_{ik}^r g_{rl} - \Gamma_{il}^r g_{kr} = 0, \quad (5.17b)$$

$$g_{li,k} - \Gamma_{kl}^r g_{ri} - \Gamma_{ki}^r g_{lr} = 0. \quad (5.17c)$$

We stress that these last three are really the same equation. Next, we add (a) and (b) and subtract (c) to obtain

$$(g_{ik,l} + g_{kl,i} - g_{li,k}) - 2\Gamma_{il}^r g_{kr} = 0. \quad (5.18)$$

In obtaining (5.18) we have made use of the symmetry of the metric and also assumed the connections are symmetric in the lower indices. To solve for the connections we multiply (5.18) by g^{kt} and contract on k to find

$$\Gamma_{il}^n = \frac{1}{2} g^{kn} (g_{ik,l} + g_{kl,i} - g_{li,k}). \quad (5.19)$$

Thus the connections are explicitly solved in terms of the metric and its derivatives. Notice their explicit symmetry in the lower indices, which we will use often. The connections defined in (5.19) are called the Christoffel connections or often the Christoffel symbols. They apply specifically to parallel displacement rather than the more general vector transplantation.

A historical note: there are also “Christoffel symbols of the first kind” used by some authors, which are defined as

$$[il, k] = \frac{1}{2} (g_{ik,l} + g_{kl,i} - g_{li,k}), \quad \Gamma_{il}^t = g^{kt} [il, k], \quad (5.20)$$

We do not use these in this book. Also, Christoffel originally used a curly bracket notation for the affine connections, but this is now seldom used (Pauli 1958; Adler 1975). In the rest of this book we will use only the connections for parallel displacement (5.19), denoted with a capital gamma, and refer to them as either Christoffel connections or simply connections.

Let us summarize properties of parallel displacement: vector transplantation using the connections defined in (5.19) gives a vector at the nearby point that is parallel to the original one in a generalized sense of parallel. Explicitly, the change in a parallel displaced vector is expressed by

$$d\xi^n + \Gamma_{li}^n dx^l \xi^i = 0, \\ \Gamma_{li}^n = \frac{1}{2} g^{kn} (g_{ik,l} + g_{kl,i} - g_{li,k}), \text{ parallel displacement.} \quad (5.21)$$

One consequence of the definition is that under parallel displacement the inner product of a vector with itself is unchanged, which means that its length remains unchanged.

Although it may appear somewhat formal at this point the idea of parallel displacement turns out to have beautiful physical and geometric meaning. It leads to definitions for the generalized straight lines called geodesics and curvature in a Riemann space. It is the central idea when we discuss covariant derivatives in tensor analysis.

It might look as if the Christoffel connections require a lot of algebra to calculate, since there are 40 of them in 4-dimensional space and $n^2(n+1)/2$ in n -dimensional space (see Exercise 5.1). Fortunately there is a shortcut method to obtain the nonzero connections using the algebra of geodesics, which we will study in Sect. 5.5 and Example 5.3.

5.4 Geodesics as Self-parallel Curves

We now know how to displace a vector to a nearby point so that it remains parallel to itself in the general sense defined in the preceding two sections. We may use this to define and study the idea of a generalized straight line or geodesic. Our definition of a geodesic stems naturally from classical Euclidean geometry and intuition. Suppose we have a curve C specified by giving the coordinates as functions of some scalar parameter p which labels points on C ,

$$\text{Curve } C: x^\mu = x^\mu(p). \quad (5.22)$$

We call C a geodesic if it is everywhere parallel to itself; this means that if we parallel displace a tangent vector along C then it remains a tangent vector.

This definition leads to a differential equation for the geodesic. Call the tangent vector $t^\alpha(p)$ at p . We parallel displace it along the curve to a nearby point labeled p' at a coordinate distance dx^α to obtain

$$t^{*\alpha}(p') = t^\alpha(p) - \Gamma_{\beta\gamma}^\alpha dx^\beta t^\gamma(p). \quad (5.23)$$

The actual tangent at p' may be obtained from that at p by a Taylor Series expansion

$$t^\alpha(p') = t^\alpha(p) + \frac{dt^\alpha}{dp} dp. \quad (5.24)$$

By our above definition of a geodesic the parallel displaced tangent vector in (5.23) is to be equal to the actual tangent vector in (5.24), so that

$$\frac{dt^\alpha}{dp} dp = -\Gamma_{\beta\gamma}^\alpha dx^\beta t^\gamma. \quad (5.25)$$

We may choose the curve parameter p to be the curve length, that is $dp = ds$, and use the normalized position derivative as an obvious tangent vector, normalized to unity,

$$t^\beta = \frac{dx^\beta}{ds}. \quad (5.26)$$

Substituting this into (5.25) we obtain a differential equation for the geodesic

$$\frac{d^2 x^\alpha}{ds^2} + \Gamma_{\beta\gamma}^\alpha \frac{dx^\beta}{ds} \frac{dx^\gamma}{ds} = 0. \quad (5.27)$$

This is termed the canonical form of the geodesic equation. Differentiation with respect to the line element s is often denoted by a dot, analogous to the time derivative in Newtonian mechanics, so the geodesic equation may be written in compact form

$$\ddot{x}^\alpha + \Gamma_{\beta\gamma}^\alpha \dot{x}^\beta \dot{x}^\gamma = 0, \quad \dot{x}^\alpha \equiv \frac{dx^\alpha}{ds}. \quad (5.28)$$

We will find this form of the equation and the notation to be useful when we consider extremum curves and some ideas of classical mechanics below.

There is a caveat to mention concerning the above analysis. In our approach to general relativity we use a signature $(1, -1, -1, -1)$ so the line element ds^2 can be positive for some curves, negative for others, and zero for others. For timelike curves, ds^2 positive, the above analysis is valid; for spacelike curves, ds^2 negative, we need merely substitute the absolute value $\sqrt{|ds^2|}$ for ds and the analysis remains valid (see Exercise 5.6). Our choice of the signature makes the arc length equal to the proper time along the trajectory of a particle. This is a convenient choice but as we discussed previously there is no universal agreement about the overall sign of the signature. We defer discussion of curves for which ds^2 is zero until later.

There is a useful and interesting property of parallel displacement along a geodesic: in a space with a positive definite metric, that is with signature $(1 \dots 1)$, the angle between any vector V and the geodesic tangent vector t may be defined as

$$\cos \theta = \frac{V^k t^j g_{jk}}{|V||t|}, \quad |V| \equiv \sqrt{V^k V^j g_{jk}}, \quad |t| \equiv \sqrt{t^k t^j g_{jk}}. \quad (5.29)$$

Under parallel displacement of a vector along a geodesic curve it is therefore obvious from the definition of parallel displacement that both the length of the displaced vector and the angle between the displaced vector and the geodesic line are unchanged.

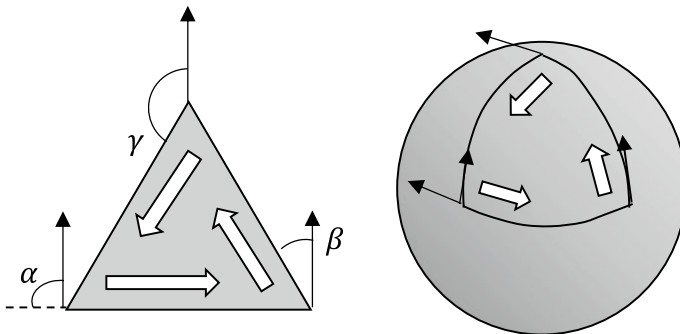


Fig. 5.4 In **a** a vector is parallel displaced around a triangle in Euclidean space. In **b** a vector is displaced around a triangle with all right angles on the surface of a sphere. The sides of both triangles are geodesics.

Example 5.1 The above constant angle property can give us insight on how parallel displacement works in flat and curved spaces. Let us parallel displace a vector around a triangle in Euclidean 2-space and also parallel displace one around a triangle on the surface of a sphere as shown in Fig. 5.4.

In flat space (a) the angle between the vector and the base of the triangle is chosen to be $\alpha = 90^\circ$ at the lower left corner; at the lower right corner the angle between the displaced vector and the right side of the triangle becomes $\beta = 30^\circ$; at the top the angle between the displaced vector and the left side of the triangle becomes $\gamma = 150^\circ$; finally at the lower left corner the displaced vector returns to its original orientation and the angle between it and the base returns to 90° . On the sphere (b) we repeat the analogous displacements around a large triangle, an octant of the sphere between the equator and the north pole. The figure makes it clear that the vector changes its orientation by 90° .

We say that the process of parallel displacing a vector is generally *not integrable*, meaning that a parallel displaced vector at a given point has an orientation that depends on the path taken to reach the point as illustrated in Fig. 5.4. Parallel displacement on a curved surface is not integrable. As we will study in Chap. 8 this is a fundamental and defining characteristic of a curved space in general.

5.5 Geodesics as Extremum Curves

The self-parallel definition of a geodesic is one of several equivalent ones. In Euclidean geometry a straight line is the shortest distance between two given points. This property can be generalized to give the following definition of a geodesic: let the curve C have length s between two fixed points; then C is a geodesic if the length s is an *extremum*, that is it is either the *shortest or longest* among all nearby curves. We will show that this definition leads to the differential equation (5.28) and is equivalent to the self-parallel definition. The extremum calculation is a problem in the calculus of variations, well-known in classical mechanics. If the reader is not familiar with such problems and the Euler-Lagrange method of solution he should first consult Appendix 2.

As before the curve C is denoted by

$$\text{Curve } C: x^\mu = x^\mu(p). \quad (5.30)$$

Here p is an invariant parameter, which may be the arc length of the curve but need not be. This is shown schematically in Fig. 5.5.

The line element along the curve and the arc length s can be written as

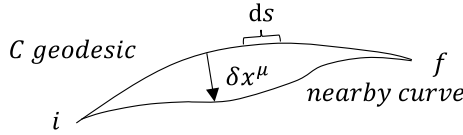


Fig. 5.5 Curve C is labeled by the invariant parameter p , and has line element ds and arc length s

$$ds^2 = g_{\alpha\beta} dx^\alpha dx^\beta = g_{\alpha\beta} \dot{x}^\alpha \dot{x}^\beta dp^2 \equiv T(x^\lambda, \dot{x}^\kappa) dp^2, \quad \dot{x}^\kappa \equiv \frac{dx^\kappa}{dp}, \quad (5.31a)$$

$$s = \int_i^f \sqrt{g_{\alpha\beta} \dot{x}^\alpha \dot{x}^\beta} dp = \int_i^f \sqrt{T(x^\lambda, \dot{x}^\kappa)} dp, \quad (5.31b)$$

where we have assumed the line element ds^2 is positive. Finding the extremum of this arc length integral is a standard problem in the calculus of variations and solvable by the Euler-Lagrange method. Indeed it is the analog of a classical mechanics problem with a Lagrangian

$$L = \sqrt{T(x^\lambda, \dot{x}^\kappa)}, \quad T(x^\lambda, \dot{x}^\kappa) \equiv g_{\alpha\beta} \dot{x}^\alpha \dot{x}^\beta. \quad (5.32)$$

The Euler-Lagrange equations for this L would give the extremum curve.

However we will do this problem in a rather subtle way to make it more useful. Most importantly our method will provide a way to get the affine connections via an elegant shortcut discussed below in Example 5.3. Instead of the square root of T let us consider any *monotonic function* F of T as the Lagrangian, and minimize the quantity

$$S = \int_i^f F(T) dp. \quad (5.33)$$

The Euler-Lagrange equations for the extremum are then

$$\frac{d}{dp} \left(\frac{\partial F}{\partial \dot{x}^\alpha} \right) - \frac{\partial F}{\partial x^\alpha} = 0 \quad \text{or} \quad \frac{d}{dp} \left(\frac{dF}{dT} \frac{\partial T}{\partial \dot{x}^\alpha} \right) - \frac{dF}{dT} \frac{\partial T}{\partial x^\alpha} = 0. \quad (5.34)$$

This equation holds along the F extremum curve. Now we choose the curve parameter p to be the curve length s ; the function T then has the constant value 1, as we see from its definition,

$$T = g_{\alpha\beta} \dot{x}^\alpha \dot{x}^\beta = g_{\alpha\beta} \frac{dx^\alpha}{ds} \frac{dx^\beta}{ds} = \frac{ds^2}{ds^2} = 1. \quad (5.35)$$

Since F is a function of only T the derivative dF/dT is a function of only T , and since T is a constant along the curve C the function dF/dT is also a constant, so we may factor it out of (5.34). This leaves

$$\frac{d}{ds} \left(\frac{\partial T}{\partial \dot{x}^\alpha} \right) - \frac{\partial T}{\partial x^\alpha} = 0. \quad (5.36)$$

That is T must obey the Euler-Lagrange equations on the extremum curve. We have thus shown algebraically that T and any monotonic function of it lead to the same extremum curve (one might expect this intuitively). This means we may study the extremum problem using not the square root of T but T itself, which is often easier.

Now we need to find the extremum of the quantity S and show that the extremum curve is the same as the geodesic curve previously defined as self-parallel. As above we choose the curve parameter to be the arc length, so the quantity to be extremized is

$$S = \int_i^f T ds = \int_i^f g_{\alpha\beta} \dot{x}^\alpha \dot{x}^\beta ds. \quad (5.37)$$

The Euler-Lagrange equations are obtained as follows

$$\begin{aligned} \frac{\partial T}{\partial \dot{x}^\lambda} &= 2g_{\lambda\alpha} \dot{x}^\alpha, \quad \frac{d}{ds} \frac{\partial T}{\partial \dot{x}^\lambda} = 2g_{\alpha\lambda} \ddot{x}^\alpha + 2g_{\alpha\lambda, \rho} \dot{x}^\alpha \dot{x}^\rho, \quad \frac{\partial T}{\partial x^\lambda} = g_{\alpha\beta, \lambda} \dot{x}^\alpha \dot{x}^\beta, \\ g_{\alpha\lambda} \ddot{x}^\alpha + g_{\alpha\lambda, \beta} \dot{x}^\alpha \dot{x}^\beta - \frac{1}{2} g_{\alpha\beta, \lambda} \dot{x}^\alpha \dot{x}^\beta &= 0. \end{aligned} \quad (5.38)$$

Next, we multiply through by $g^{\mu\lambda}$ and juggle indices using the symmetry of the metric, to obtain

$$\ddot{x}^\mu + \frac{1}{2} g^{\mu\lambda} (g_{\lambda\beta, \alpha} + g_{\alpha\lambda, \beta} - g_{\alpha\beta, \lambda}) \dot{x}^\alpha \dot{x}^\beta = 0, \quad \ddot{x}^\mu + \Gamma_{\alpha\beta}^\mu \dot{x}^\alpha \dot{x}^\beta = 0. \quad (5.39)$$

Thus the extremum curve is a geodesic since it satisfies the same differential equation (5.28) as we previously obtained for the geodesic.

Several features of this result are worth noting. We specialized to the arc length as the curve parameter, but it is evident from the geodesic equation that any constant *multiple* of the arc length will give the same equation, that is dp proportional to ds . Also it is obvious that our approach cannot be used if the geodesic is a null curve, one along which the line element is zero, $ds = 0$. We will consider this special case later. Finally note that in Euclidean space the interesting geodesics are the shortest curves between points, while in the Minkowski space of special relativity they are the longest curves between points, as we discussed in Sect. 3.4.

Example 5.2 It is illustrative to study a simple example of the extremum approach—to get the geodesics in Euclidean 2-space. Using polar coordinates

(ρ, φ) the line element and the corresponding T function are

$$ds^2 = d\rho^2 + \rho^2 d\varphi^2, \quad T = \dot{\rho}^2 + \rho^2 \dot{\varphi}^2. \quad (5.40)$$

From this we may get the Euler-Lagrange equations. For the ρ equation

$$\begin{aligned} \frac{\partial T}{\partial \dot{\rho}} &= 2\dot{\rho}, \quad \frac{d}{ds} \frac{\partial T}{\partial \dot{\rho}} = 2\ddot{\rho}, \quad \frac{\partial T}{\partial \rho} = 2\rho \dot{\varphi}^2, \\ \ddot{\rho} - \rho \dot{\varphi}^2 &= 0. \end{aligned} \quad (5.41)$$

For the φ equation

$$\frac{\partial T}{\partial \dot{\varphi}} = 2\rho^2 \dot{\varphi}, \quad \frac{d}{ds} \frac{\partial T}{\partial \dot{\varphi}} = 2\rho^2 \ddot{\varphi} + 4\rho \dot{\rho} \dot{\varphi}, \quad \frac{\partial T}{\partial \varphi} = 0, \quad (5.42)$$

$$\rho^2 \ddot{\varphi} + 2\rho \dot{\rho} \dot{\varphi} = 0, \quad \text{so} \quad \rho^2 \dot{\varphi} = \text{const.}$$

You should check that a ray, constant φ , is a solution to the last two equations.

There is a beautiful practical use for the two approaches to geodesics we have just worked through. It is apparent that the connections could be tedious to calculate from their definition since there may be a large number of them, 40 in the four dimensions of relativity. However the relation we have just worked out between the Euler-Lagrange equations and the geodesic equation provides a simple useful shortcut. The geodesic equations may be written in both the Euler-Lagrange form (5.36), which is often easy, and also the standard canonical form (5.39) containing the connections. We need only compare the two to pick out the nonzero connections. Example 5.3 illustrates this.

Example 5.3 To show how this shortcut works we will apply it to the case of polar coordinates in Euclidean 2-space that we worked with above. We obtained the Euler-Lagrange equations in the previous example, so we compare them with the canonical form. For the $x^1 = \rho$ equation in (5.42) and the canonical form in (5.39)

$$\ddot{\rho} - \rho \dot{\varphi}^2 \Leftrightarrow 0 - \ddot{\rho} + \Gamma_{ij}^1 \dot{x}^i \dot{x}^j = 0. \quad (5.43)$$

From this it is apparent that only one of the connections with an upper index 1 is nonzero,

$$\Gamma_{22}^1 = -\rho. \quad (5.44)$$

Similarly for the $x^2 = \varphi$ equation

$$\ddot{\varphi} + \frac{2}{\rho} \dot{\rho} \dot{\varphi} \Leftrightarrow 0 - \ddot{\varphi} + \Gamma_{ij}^2 \dot{x}^i \dot{x}^j = 0. \quad (5.45)$$

From this there are only two equal nonzero connections with an upper index 2,

$$\Gamma_{21}^2 = \Gamma_{12}^2 = \frac{1}{\rho}. \quad (5.46)$$

The ease of the technique is apparent, especially so since the metric is diagonal and most of the connections are zero. It is often a large labor-saving technique.

In the rest of this book we will make frequent use of the technique in Example 5.3 for calculating the connections.

5.6 Affine Connections, Abstract View

Let us see how we may motivate and interpret the coefficients of affine connection using the abstract view introduced in Sect. 4.3. Recall that a vector may be expanded in a coordinate basis, that is vectors aligned along the coordinate axes, according to

$$\vec{V} = V^j \vec{e}_j, \quad \vec{e}_j = \text{coordinate basis}. \quad (5.47)$$

If we think of moving the vector to a nearby point it will change due to a change in its components and also a change in the basis vectors,

$$d\vec{V} = \vec{e}_i dV^i + V^j d\vec{e}_j. \quad (5.48)$$

As we discussed previously the vector spaces associated with different points in a Riemann space are ab initio independent. As such it is necessary to postulate a way to relate them. This leads to the idea of vector transplantation and the specific version of transplantation called parallel displacement that we discussed in Sects. 5.1 and 5.3. We can think of this in the present abstract view as giving an effective change in the coordinate basis, which we assume is a bilinear expression in the basis vectors and the coordinate displacement; it is a rather compelling assumption. That is we postulate

$$d\vec{e}_j = \Gamma_{kj}^i (\vec{e}_i dx^k). \quad (5.49)$$

The coefficients in the expansion will of course be identified as the affine connections. They can then be obtained explicitly by asking that inner products be unchanged when parallel displaced just as we did in Sect. 5.3. Thereby the affine space plus a geometric or physical demand becomes the Riemann space we use in general relativity theory.

Now we substitute the basis change (5.49) in the vector change (5.48) and obtain

$$d\vec{V} = (dV^i + \Gamma_{kj}^i V^j dx^k) \vec{e}_i. \quad (5.50)$$

The condition that the vector components change according to the law (5.5) of vector transplantation thus corresponds to $d\vec{V} = 0$. We have obtained an interpretation for the affine connections as being related to the change in the coordinate basis vectors (5.49).

Note that the defining expression (5.50) for the connections does *not* imply that they must be symmetric in the lower indices; the same is true in the component view for (5.5).

In the case of flat space the derivatives of the basis vectors may be calculated explicitly and we could rewrite (5.49) in terms of those derivatives. This leads to an explicit expression for the connections in that special case,

$$d\vec{e}_j = \vec{e}_{j,k} dx^k \equiv (\Gamma_{kj}^i \vec{e}_i) dx^k, \quad \text{so} \quad \Gamma_{kj}^n = (\vec{e}_{j,k} \cdot \vec{e}_i) g^{ni}. \quad (5.51)$$

Geodesics in terms of vector transplantation fit naturally into the present scheme. Suppose as usual that we have a curve C with arc length s . Then any vector $\vec{V}$ defined along the curve will change according to (5.50) and have a derivative along the curve given by

$$\frac{d\vec{V}}{ds} = \left(\frac{dV^i}{ds} + \Gamma_{kj}^i V^j \frac{dx^k}{ds} \right) \vec{e}_i. \quad (5.52)$$

Apply this relation now to a tangent vector to the curve, which we may take to be

$$\vec{\tau} = \frac{dx^k}{ds} \vec{e}_k. \quad (5.53)$$

Then the derivative of the tangent vector along the curve is

$$\frac{d\vec{\tau}}{ds} = \left(\frac{d^2 x^i}{ds^2} + \Gamma_{kj}^i \frac{dx^j}{ds} \frac{dx^k}{ds} \right) \vec{e}_i. \quad (5.54)$$

Thus if we define a geodesic as having a constant tangent vector we find that the curve is the same as the geodesic curve we defined in Sect. 5.4,

$$\dot{x}^i + \Gamma_{kj}^i \dot{x}^k \dot{x}^j = 0, \quad \dot{x}^k = \frac{dx^k}{ds}, \quad \text{geodesic equation.} \quad (5.55)$$

The above treatment of the geodesic curve has several interesting features that should be noted. First, the connections and the metric may be treated independently; in the general case it is not necessary to have a relation between the two. Any affine space can thus admit geodesic curves. Secondly, the connections according to (5.49) need not be symmetric in the lower indices, unlike the Christoffel connections, so torsion is admissible; that is, the connections may have antisymmetric parts. However, according to (5.55) the anti-symmetric part of the connection cancels out of the geodesic equation. We will see later in Chap. 7 that the geodesic equation determines the motion of bodies in relativity theory, so torsion would have no effect on such motion. The physical relevance of torsion in the context of relativity theory is indeed not obvious (Trautman 2006).

Appendix 1: A Special Coordinate System

Recall that in Chap. 4 we stated the Signature Theorem, that at any point P there exists a special coordinate system in which the metric is diagonal and has diagonal elements equal to 1 or -1 or 0. The special system may be reached by a linear transformation. This form of the metric is called the Cayley-Sylvester canonical form. We proved the theorem for the case of two dimensions in Appendix 4.1 (Perlis 1952).

In this chapter we obtained another special coordinate system, the geodesic system, in which the affine connections vanish at any given point P . If the connections are zero, then from the definition of the Christoffel connections (5.19) this clearly means that the first derivatives of the metric must also be zero. We can in fact combine these transformations and for any given point P find a coordinate system in which the metric has the Cayley-Sylvester canonical form and also has vanishing first derivatives and thus vanishing connections. To do this we merely apply the two transformations together with the point P taken to be the origin,

$$\bar{x}^j = L_k^j x^k + \frac{1}{2} A_{jl}^i (L_n^j x^n) (L_m^l x^m). \quad (5.56)$$

The L array makes the transformation to the system in which the metric has the Cayley-Sylvester canonical form, and the A array specifies the transformation to the geodesic system. The coordinate system thus obtained is very special: the axes are orthogonal, the metric is Lorentz, and the connections vanish, so physics is locally much like that of special relativity, but of course only in a vanishingly small region near P .

Appendix 2: The Extremum Problem and the Euler-Lagrange Equations

For completeness we briefly review one of the most important problems in the calculus of variations, one which is familiar to most physicists from the Lagrangian formulation of classical mechanics (Goldstein 1980). The Lagrangian is assumed to be a given function of the coordinates and generalized velocities, $L(x^\lambda, \dot{x}^\alpha)$. A quantity S called the action is then defined as the integral of the Lagrangian along some curve from a fixed initial point i to a fixed final point f ,

$$S = \int_i^f L(x^\lambda, \dot{x}^\alpha) dp, \quad \dot{x}^\alpha \equiv \frac{dx^\alpha}{dp}. \quad (5.57)$$

That is, the action is a *functional* of the Lagrangian. The Euler-Lagrange method of extremizing the action is to calculate the variation in S as the path $x^\mu(p)$ is varied by a small amount $\delta x^\mu(p)$ as shown in Fig. 5.5; the extremum path is characterized by the vanishing of the variation, precisely analogous to the vanishing of a derivative of a function at its extremum. The variation in S is calculated in a straight-forward way as follows,

$$\begin{aligned} \delta S &= \int_i^f \left[\frac{\partial L}{\partial x^\alpha} \delta x^\alpha + \frac{\partial L}{\partial \dot{x}^\alpha} \delta \dot{x}^\alpha \right] dp \\ &= \int_i^f \left[\frac{\partial L}{\partial x^\alpha} \delta x^\alpha + \frac{d}{dp} \left(\frac{\partial L}{\partial \dot{x}^\alpha} \delta x^\alpha \right) - \delta x^\alpha \frac{d}{dp} \left(\frac{\partial L}{\partial \dot{x}^\alpha} \right) \right] dp \\ &= \int_i^f \left[\frac{\partial L}{\partial x^\alpha} - \frac{d}{dp} \left(\frac{\partial L}{\partial \dot{x}^\alpha} \right) \right] \delta x^\alpha dp + \left(\frac{\partial L}{\partial \dot{x}^\alpha} \delta x^\alpha \right)_i^f, \end{aligned} \quad (5.58)$$

where we have integrated by parts and used $\delta \dot{x}^\alpha = d(\delta x^\alpha)/dp$. Since we consider only paths between fixed endpoints the last term in the last line above is zero. Since we consider any small variation δx^α the bracket in the integral must be identically zero, so we conclude

$$\frac{d}{dp} \left(\frac{\partial L}{\partial \dot{x}^\alpha} \right) - \frac{\partial L}{\partial x^\alpha} = 0. \quad (5.59)$$

These differential equations are called the Euler-Lagrange equations, and yield a curve for which the action is extremum.

It would be hard to exaggerate the utility of the Euler-Lagrange type of analysis and the action concept in classical and quantum mechanics, classical and quantum field theory, and essentially all of physics.

Appendix 3: Christoffel Connections as Fictitious Forces

The Christoffel connections are actually familiar objects in classical mechanics, but they are seldom identified as such explicitly or seen from the geometrical point of view. They give rise to the well-known fictitious forces encountered in non-cartesian coordinate systems, rotating systems being a favorite example. To illustrate how this works we will study the motion of a particle in a potential in 3-dimensional space with a general coordinate system using the Lagrangian formulation of classical mechanics. The manipulations are similar to those used in the preceding appendix and for discussing geodesics in the text.

Let the particle have a trajectory in three dimensions, with the position is given as a function of absolute (invariant) time by $x^j(t)$ in some coordinate system. Along this trajectory the line element represents the Euclidean distance

$$ds^2 = g_{ij}dx^i dx^j. \quad (5.60)$$

Thus we may write the square of the velocity as

$$v^2 = g_{ij}\dot{x}^i \dot{x}^j, \quad \dot{x}^i \equiv \frac{dx^i}{dt}. \quad (5.61)$$

For a particle moving in a potential field the Lagrangian is generally taken to be the kinetic energy minus the potential energy,

$$L = \frac{m}{2}v^2 - V(x^k) = \frac{m}{2}g_{ij}\dot{x}^i \dot{x}^j - V(x^k). \quad (5.62)$$

Note the similarity of this to the function T which we used in discussing geodesics. Lagrangian mechanics is based on the postulate that the action, the integral of L , is extremized for the correct trajectory. That is

$$\delta S = 0, \quad S = \int_i^f L dt = \int_i^f \left[\frac{m}{2}g_{ij}\dot{x}^i \dot{x}^j - V(x^k) \right] dt. \quad (5.63)$$

Extremizing the action we are led to the Euler-Lagrange equations as in our derivation of the geodesic equation, but now we also have a potential energy term. The Euler-Lagrange equations are obtained as usual, and are,

$$\begin{aligned}
\frac{\partial L}{\partial \dot{x}^i} &= m g_{ij} \dot{x}^j, & \frac{d}{dt} \frac{\partial L}{\partial \dot{x}^i} &= m (g_{ij} \ddot{x}^j + g_{ij,k} \dot{x}^j \dot{x}^k), \\
\frac{\partial L}{\partial x^i} &= \frac{m}{2} g_{ij,k} \dot{x}^j \dot{x}^k - \frac{\partial V}{\partial x^i}, \\
m \left(g_{ij} \ddot{x}^j + g_{ij,k} \dot{x}^j \dot{x}^k - \frac{1}{2} g_{jk,i} \dot{x}^j \dot{x}^k \right) + \frac{\partial V}{\partial x^i} &= 0.
\end{aligned} \tag{5.64}$$

Finally we multiply by g^{ki} and rearrange indices to obtain

$$m [\ddot{x}^k + \Gamma_{ji}^k \dot{x}^j \dot{x}^i] = -g^{ki} \frac{\partial V}{\partial x^i} \equiv F^k. \tag{5.65}$$

This is essentially Newton's second law in an arbitrary coordinate system. The force is defined as usual as the negative of the gradient of the potential energy, and is a contravariant vector. In this formulation we see that the second term in the bracket plays the same role as a force in producing the acceleration $\ddot{x}^k$; such forces are called *fictitious* because they do not occur in a Cartesian coordinate system and may be transformed away. Indeed, the Weyl Theorem, Theorem 4, shows explicitly how this is done. Notice that one of the characteristics of a fictitious force is that it is proportional to the mass, a fact that has fundamental importance in the physics of gravity.

Finally we point out that if the force vanishes then the particle follows a geodesic, as apparent from (5.65). This is further motivation for interpreting a geodesic as a generalized straight line.

A word of caution is in order concerning the word “fictitious” for the forces represented by Christoffel connections in (5.65). These forces cause acceleration like any other force, and are thus no less real. In particular they are quite as real as gravity which can also be transformed away as we will see in Part III. Because of this, many physicists do not approve of the word fictitious, but the name is now entrenched and we continue to use it with this proviso.

Exercises

- 5.1 How many independent affine connections are there in 2, 3, 4 and n dimensions if they are assumed to be symmetric in the lower indices? What if they are not symmetric?
- 5.2 What are the Christoffel connections for Euclidean 2-space, Euclidean 3-space, and Euclidean n -space with Cartesian coordinates? What of a space and coordinate system with a more general but constant metric field? (This is as easy as it sounds!)
- 5.3 Using their definition work out the Christoffel connections for the simple case of Euclidean 2-space using polar coordinates.
- 5.4 Repeat Exercise 5.3 for a non-flat surface with metric

$$ds^2 = f(\rho)^2 d\rho^2 + \rho^2 d\varphi^2,$$

where $f(\rho)$ is a smooth function of ρ . Obtain the geodesic equations and show that rays $\varphi = \text{const.}$ are geodesics. This should also be intuitively obvious.

- 5.5 Go through the derivation of the geodesic equation for a curve with negative line element as mentioned briefly in the text, leading to (5.28).
- 5.6 All humans to date have lived on or near the surface of the spherical earth, so it is a good idea to study connections and geodesics for a spherical surface. Indeed the word geodesic derives from “dividing the earth” in Greek. Write down the metric in terms of spherical coordinates with constant radius. From the metric write down the function T defined in (5.32) to be used as a Lagrangian. From this T write down the Euler-Lagrange equations which describe a geodesic on the surface. For a sphere these are also called great circles.
- 5.7 Continue working on the sphere. From the Euler-Lagrange equations in Exercise 5.6 show that the equator and longitude lines are geodesics but latitude lines are not. This should also be obvious.
- 5.8 How many affine connections are there on the spherical surface? Compare the Euler-Lagrange equations with the geodesic equations in standard form (5.28) and identify the affine connections using the procedure we used in Example 5.3.
- 5.9 Use classical Lagrangian mechanics to study the motion of a particle in a 2-dimensional plane, with a central potential energy field, using polar coordinates; that is, write down the Euler-Lagrange equations. For the case of zero force note that one obtains the equation of a simple straight line. Is the physical meaning clear?
- 5.10 The Gauss-Bonnet Theorem relates the angle of rotation of a vector parallel displaced along geodesics on a closed curve to the area enclosed by the curve. Find a reference on this theorem and verify it for the sphere shown in Fig. 5.4 in which the closed curve is a triangle with all right angles. This theorem provides one way to define curvature.

Chapter 6

Tensor Analysis



Abstract The ideas of classical vector analysis in Euclidian space generalize naturally to Riemann space. Affine connections are the key to this generalization. Moreover much of classical vector analysis becomes more clear and simple; the divergence and Laplacian are prime examples.

6.1 Covariant Derivatives, Component View

We know that the derivative of a scalar function is a covariant vector from Chap. 4, so it has well-defined tensor transformation properties. The derivative of a vector field is not so simple however. In the preceding chapter we learned how to displace a vector parallel to itself in an elegant and general way, and we now use the parallel displacement concept to form a new kind of derivative. Consider the vector field $W^i(x^j)$. In going from a point in space x^j to a nearby point $x^j + dx^j$ the field changes according to

$$W^i(x^l + dx^l) = W^i(x^l) + W^i_{,k}(x^l)dx^k. \quad (6.1)$$

If it were parallel displaced to the new point it would change according to

$$W^{*i}(x^l + dx^l) = W^i(x^l) - \Gamma^i_{kj}(x^l)dx^k. \quad (6.2)$$

If the vector field were constant, in the sense of being parallel to itself, then these two would be equal. Thus we may think of the *relevant* change in the field as the *difference* between the actual value of the field at $x^j + dx^j$ and the value it would have if parallel displaced there from x^j . This is shown in Fig. 6.1.

Accordingly we define the covariant derivative of W^i in terms of this difference via

$$\begin{aligned} & W^i(x^l + dx^l) - W^{*i}(x^l + dx^l) \\ &= [W^i_{,k}(x^l) + \Gamma^i_{kj}(x^l)W^j(x^l)]dx^k = W^i_{;k}dx^k, \end{aligned} \quad (6.3a)$$

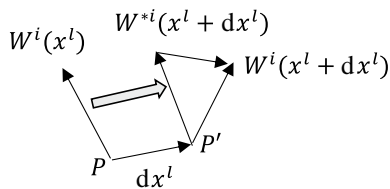


Fig. 6.1 The relevant change considered for the covariant derivative is the difference between the vector field at the new point P' and the vector *parallel displaced* there from the original point P

$$W^i{}_{;k} \equiv W^i{}_{,k} + \Gamma^i_{kj} W^j. \quad (6.3b)$$

We used a comma before to denote the ordinary derivative and now we use a semi-colon to denote the covariant derivative. In the special case in which a vector field has a zero covariant derivative in a small region the field is parallel to itself in that region, and we think of it as constant in a generalized sense.

We have defined the covariant derivative in a rather natural way, but we have not yet justified the name covariant; the justification is in the following theorem:

Theorem 1 *The covariant derivative of a contravariant vectorfield is a (1,1) tensor. Moreover It will be clear from the proof that the ordinary derivative is not a tensor. The proof is straight-forward because we know how vectors, ordinary derivatives, and connections transform. There is just a bit of index juggling algebra involved. From the vector transformation law in (4.14) and the chain rule we first calculate the transformation of the ordinary derivative,*

$$\bar{W}^i = \frac{\partial \bar{x}^i}{\partial x^l} W^l, \quad \frac{\partial}{\partial \bar{x}^k} = \frac{\partial x^j}{\partial \bar{x}^k} \frac{\partial}{\partial x^j}, \quad (6.4)$$

thus

$$\frac{\partial \bar{W}^i}{\partial \bar{x}^k} = \left(\frac{\partial x^j}{\partial \bar{x}^k} \frac{\partial}{\partial x^j} \right) \left[\frac{\partial \bar{x}^i}{\partial x^l} W^l \right] = \frac{\partial x^j}{\partial \bar{x}^k} \frac{\partial \bar{x}^i}{\partial x^l} \frac{\partial W^l}{\partial x^j} + \frac{\partial x^j}{\partial \bar{x}^k} \frac{\partial^2 \bar{x}^i}{\partial x^j \partial x^l} W^l.$$

This is not the transformation law of a tensor. From the transformation of the connections (5.11) we may calculate the second term in the covariant derivative in the barred frame,

$$\begin{aligned} \bar{\Gamma}^i_{jk} \bar{W}^j &= \left[\frac{\partial \bar{x}^i}{\partial x^l} \frac{\partial x^p}{\partial \bar{x}^j} \frac{\partial x^q}{\partial \bar{x}^k} \Gamma^l_{pq} - \frac{\partial^2 \bar{x}^i}{\partial x^m \partial x^l} \frac{\partial x^m}{\partial \bar{x}^k} \frac{\partial x^l}{\partial \bar{x}^j} \right] \frac{\partial \bar{x}^j}{\partial x^n} W^n \\ &= \frac{\partial \bar{x}^i}{\partial x^l} \frac{\partial x^j}{\partial \bar{x}^k} \Gamma^l_{jn} W^n - \frac{\partial^2 \bar{x}^i}{\partial x^j \partial x^l} \frac{\partial x^j}{\partial \bar{x}^k} W^l. \end{aligned} \quad (6.5)$$

As usual we have made liberal use of relabeling summation indices. Finally we combine the last two equations to obtain the complete transformation, written in three equivalent ways,

$$\frac{\partial \bar{W}^i}{\partial \bar{x}^k} + \bar{\Gamma}_{kj}^i \bar{W}^j = \frac{\partial \bar{x}^i}{\partial x^l} \frac{\partial x^j}{\partial \bar{x}^k} \left[\frac{\partial W^l}{\partial x^j} + \Gamma_{jn}^l W^n \right], \quad (6.6a)$$

$$\bar{W}_{;k}^i + \bar{\Gamma}_{kj}^i \bar{W}^j = \frac{\partial \bar{x}^i}{\partial x^l} \frac{\partial x^j}{\partial \bar{x}^k} [W_{;j}^l + \Gamma_{jn}^l W^n], \quad (6.6b)$$

$$\bar{W}_{;k}^i = \frac{\partial \bar{x}^i}{\partial x^l} \frac{\partial x^j}{\partial \bar{x}^k} W_{;j}^l. \quad (6.6c)$$

The second derivative term has magically cancelled out. The transformation law is that of a second rank tensor, once contravariant and once covariant or (1,1).

This theorem makes it clear that the covariant derivative is the natural generalization of the ordinary derivative since it is a tensor and reduces to the ordinary derivative in flat space with Cartesian coordinates.

We now have derivatives for a scalar field and for a vector field that are tensors. From these we can infer unique definitions for derivatives of other tensors. To obtain a definition for the covariant derivative of a covariant vector field we use what we already know about the derivatives of the scalar and the contravariant vector fields; we demand that the product rule (or Leibniz rule) for ordinary derivatives hold also for the covariant derivative. Thus both the ordinary and covariant derivative of the scalar field $W^k V_k$ should obey the product rule. This gives

$$\begin{aligned} (W^k V_k)_{;l} &= W^k_{;l} V_k + W^k V_{k;l} \quad \text{ordinary derivative,} \\ (W^k V_k)_{;l} &= W^k_{;l} V_k + W^k V_{k;l} \quad \text{covariant derivative.} \end{aligned} \quad (6.7)$$

But for the scalar inner product the ordinary and covariant derivatives are the same, so

$$W^k V_{k;l} = W^k V_{k,l} + W^k_{;l} V_k - W^k_{;l} V_k = W^k [V_{k,l} - \Gamma_{kl}^n V_n]. \quad (6.8)$$

Since W^k can be any vector we see that for consistency the covariant derivative must be defined as

$$V_{k;l} = V_{k,l} - \Gamma_{kl}^n V_n. \quad (6.9)$$

We have now obtained consistent definitions for the covariant derivative of scalar fields and contravariant and covariant vector fields. Using these and the product rule we may infer definitions and properties for the covariant derivative of any (M, N) tensor field. For example the covariant derivative of a second rank tensor must be the same as that for the direct product of vectors, both for consistency and because any

tensor may be written as the sum of such products, as we have previously shown. We will work out two examples and see that the general case becomes evident. First consider the (2,0) tensor $T^{ab} = U^a V^b$. Imposing the product rule for the covariant derivative we see that

$$\begin{aligned} T^{ab}{}_{;c} &= (U^a V^b)_{;c} = U^a{}_{;c} V^b + U^a V^b{}_{;c} \\ &= U^a{}_{,c} V^b + U^a V^b{}_{,c} + \Gamma_{cd}^a U^d V^b + \Gamma_{cd}^b U^a V^d \\ &= T^{ab}{}_{,c} + \Gamma_{cd}^a T^{db} + \Gamma_{cd}^b T^{ad}. \end{aligned} \quad (6.10)$$

Similarly we may repeat the procedure for a mixed (1,1) tensor $M^a{}_b = W^a A_b$.

$$\begin{aligned} M^a{}_{b;c} &= (W^a A_b)_{;c} = W^a{}_{;c} A_b + W^a A_{b;c} \\ &= W^a{}_{,c} A_b + W^a A_{b,c} + \Gamma_{cd}^a W^d A_b - \Gamma_{cb}^d W^a A_d \\ &= M^a{}_{b,c} + \Gamma_{cd}^a M^d{}_b - \Gamma_{cb}^d M^a{}_d. \end{aligned} \quad (6.11)$$

The general case is evident from these two examples: the covariant derivative is the ordinary derivative plus a connection term for each upper index and minus a connection term for each lower index. After a little practice the index placement becomes easy to remember.

The metric tensor is a very special tensor, and its covariant derivative is particularly interesting and important.

Theorem 2 (Ricci Theorem) *The covariant derivative of the metric tensor is zero.* The covariant derivative is easy to calculate from the definition just given and the definition of the connection,

$$\begin{aligned} g_{\mu\nu;\lambda} &= g_{\mu\nu,\lambda} - \Gamma_{\nu\lambda}^\alpha g_{\alpha\mu} - \Gamma_{\mu\lambda}^\alpha g_{\nu\alpha} \\ &= g_{\mu\nu,\lambda} - \frac{1}{2} g^{\alpha\tau} (g_{\lambda\tau,v} + g_{v\tau,\lambda} - g_{v\lambda,\tau}) g_{\alpha\mu} \\ &\quad - \frac{1}{2} g^{\alpha\tau} (g_{\lambda\tau,\mu} + g_{\mu\tau,\lambda} - g_{\mu\lambda,\tau}) g_{\alpha\nu} \\ &= g_{\mu\nu,\lambda} - \frac{1}{2} (g_{\lambda\mu,v} + g_{v\mu,\lambda} - g_{v\lambda,\mu}) \\ &\quad - \frac{1}{2} (g_{\lambda\nu,\mu} + g_{\mu\nu,\lambda} - g_{\mu\lambda,\nu}) = 0. \end{aligned} \quad (6.12)$$

The Ricci Theorem is thus quite easy to prove, and it is very important for consistency in the tensor derivative notation. For example, given a covariant derivative of a vector like $V^\alpha{}_{;\tau}$ there are two different things that we might mean by lowering an index to form $V_{\beta;\tau}$. These are

$$V_{\beta;\tau} = g_{\beta\alpha} (V^\alpha{}_{;\tau}) \quad \text{or} \quad V_{\beta;\tau} = (g_{\beta\alpha} V^\alpha)_{;\tau}. \quad (6.13)$$

Because of the Ricci Theorem these are the same, and there is in fact no ambiguity in the notation. Another way to say the same thing is that the operations of raising and lowering indices commutes with the operation of taking a covariant derivative.

It is also interesting to note a rather obvious converse of the Ricci Theorem: If the covariant derivative of the metric tensor is zero then the connections are the Christoffel connections. This follows because in the covariant derivative relation (6.12) the first line is the same as (5.17), which leads to the Christoffel connections in (5.19). Thus the Christoffel connections are dictated by the demands that they be symmetric and the covariant derivative of the metric be zero.

6.2 Covariant Derivatives, Abstract View

As before in Sect. 5.6 we now consider vectors as invariant abstract objects that may be expanded in a basis, conveniently taken to be a coordinate basis. Then the vector and its change in moving to a nearby point are, as discussed in Sect. 5.6,

$$\vec{V} = V^j \vec{e}_j, \quad d\vec{V} = (dV^i + \Gamma_{kj}^i V^j dx^k) \vec{e}_i, \quad \Gamma_{kj}^i = \text{affine connections.} \quad (6.14)$$

For a field of vectors $V^j = V^j(x^k)$ we thus have the change

$$\begin{aligned} d\vec{V} &= \left(\frac{\partial V^i}{\partial x^k} dx^k + \Gamma_{kj}^i V^j dx^k \right) \vec{e}_i = \left(\frac{\partial V^i}{\partial x^k} + \Gamma_{kj}^i V^j \right) dx^k \vec{e}_i \\ &= (V^i{}_{;k} dx^k) \vec{e}_i. \end{aligned} \quad (6.15)$$

This defines the coefficient array $V^i{}_{;k}$ as determining the change in the vectot. Both sides of (6.15) are invariant abstract vectors, so the last object in parentheses is the i th component of the change in the vector. By the quotient theorem the array $V^i{}_{;k}$ then forms the components of a (1,1) tensor, as we have already discussed in terms of components in Sect. 6.1. In terms of the basis vectors and forms we may write that tensor as

$$\nabla \vec{V} = V^i{}_{;k} (\vec{e} \otimes \tilde{dx}^k) \quad \text{Covariant tensor derivative.} \quad (6.16)$$

The various component arrays that we discussed in Sect. 6.1 now emerge as coefficients in tensor relations just as happened with tensors in general in Sect. 4.4. The tensor character of the covariant derivative is made particularly clear in this approach whereas in the component approach it required a bit of algebra to verify its transformation as a tensor.

Consider next the derivative of a vector field along some given curve parametrized as usual by the arc length s . From (6.15) we may define the derivative as

$$\frac{d\vec{V}}{ds} = \left(V^i{}_{;k} \frac{dx^k}{ds} \right) \vec{e}_i = (V^i{}_{;k} t^k) \vec{e}_i, \quad t^k \equiv \frac{dx^k}{ds}. \quad (6.17)$$

Here t^k are the components of the tangent vector to the curve. The object in the last parentheses is then the component array of the curve derivative of the vector.

The last expression for the curve derivative may also be written in an informative canonical form. First note that the components t^k may be expressed in terms of the abstract vector $\vec{t}$ as we see from the relations

$$\vec{t} = t^n \vec{e}_n, \quad \text{so} \quad \tilde{dx}^k(\vec{t}) = t^n \tilde{dx}^k(\vec{e}_n) = t^n \delta_n^k = t^k. \quad (6.18)$$

We substitute this for t^k into (6.17) and find

$$\frac{d\vec{V}}{ds} = V^i{}_{;k} \vec{e}_i \tilde{dx}^k(\vec{t}) = V^i{}_{;k} (\vec{e}_i \otimes \tilde{dx}^k)(-, \vec{t}). \quad (6.19)$$

This displays the curve derivative in the direction $\vec{t}$ in terms of the basis vectors and forms.

Having obtained the covariant derivative of a vector we see it is fairly obvious how to infer the necessary definition for the covariant derivative of a general tensor; the logic is much the same as used for (6.10) in Sect. 6.1. We consider the special case of a (2,0) tensor which is the direct product of two vectors,

$$T = \vec{V} \otimes \vec{W} = (V^i \vec{e}_i) \otimes (W^n \vec{e}_n) = V^i W^n (\vec{e}_i \otimes \vec{e}_n). \quad (6.20)$$

We then impose the product or Leibniz rule for derivatives and after some algebra find a relation analogous to (6.16),

$$\begin{aligned} \nabla T &= \nabla \vec{V} \otimes \vec{W} + \vec{V} \otimes \nabla \vec{W} = (V^i{}_{;k} W^n + V^i W^n{}_{;k}) (\vec{e}_i \otimes \vec{e}_n \otimes \tilde{dx}^k) \\ &= (T^{in}{}_{;k}) (\vec{e}_i \otimes \vec{e}_n \otimes \tilde{dx}^k). \end{aligned} \quad (6.21)$$

From this it is clear that the covariant derivative of any tensor must be given in terms of its components and the basis by

$$\nabla T = (T^{i\dots j\dots k}{}_{;l}) (\vec{e}_i \otimes \dots \otimes \tilde{dx}^j \otimes \dots \otimes \tilde{dx}^k). \quad (6.22)$$

That is, given the component array for the covariant derivative discussed in Sect. 6.1 the covariant derivative of the abstract tensor obeys the same sort of equation as (4.61).

The special case of the covariant derivative of the metric tensor is worth mentioning due to its importance. We have

$$\nabla g = g_{ij;k} (\tilde{dx}^i \otimes \tilde{dx}^j \otimes \tilde{dx}^k). \quad (6.23)$$

This is zero according to the Ricci Theorem of Sect. 6.1.

In brief summary the equations and various component arrays in the preceding Sect. 6.1 emerge in this section as coefficients in the abstract approach, just as happened with tensors in general in Chap. 5.

6.3 The Divergence and Laplacian

In elementary vector calculus with Cartesian coordinates the divergence of a vector field is defined as

$$\operatorname{div} \vec{B} = \nabla \cdot \vec{B} = \frac{dB_x}{dx} + \frac{dB_y}{dy} + \frac{dB_z}{dz} = B^i{}_{,i}, \quad \text{divergence.} \quad (6.24)$$

The obvious covariant generalization of this is the contracted covariant derivative,

$$B^i{}_{;i} = B^i{}_{,i} + \Gamma^k_{kl} B^l. \quad (6.25)$$

This may be simplified into a form which contains no connections and is thus easy to deal with. The contracted connection is

$$\Gamma^k_{kl} = \frac{1}{2} g^{kn} (g_{nk,l} + g_{ln,k} - g_{kl,n}) = \frac{1}{2} g^{kn} g_{nk,l}, \quad (6.26)$$

where we have used the symmetry of the metric to cancel the second and third terms.

At this point we digress to recall some properties of matrices and determinants, referred specifically to the metric tensor treated as a matrix. The inverse of the metric, g^{ik} , may be calculated as $g^{ik} = \Delta^{ik}/|g|$ where $|g|$ is the determinant and Δ^{ki} is the *cofactor* matrix; the cofactor is found by crossing out the i row, and taking the determinant with a sign $(-1)^{i+k}$. The determinant may be similarly expressed in terms of the cofactor: choose a row, say $i = 3$, and the determinant is $|g| = g_{3k} \Delta^{3k}$. This is often referred to as expansion in minors. From the above relations we see that

$$\frac{\partial |g|}{\partial g_{jk}} = \Delta^{jk} = |g| g^{jk}, \quad \text{so} \quad g^{jk} = \frac{1}{|g|} \frac{\partial |g|}{\partial g_{jk}}. \quad (6.27)$$

Now we return to the expression for the contracted connection in (6.26) and substitute the above to obtain several alternative ways to write it

$$\begin{aligned} \Gamma^k_{kl} &= \frac{1}{2} g^{kt} g_{kt,l} = \frac{1}{2} \left(\frac{1}{|g|} \frac{\partial |g|}{\partial g_{kt}} \right) \frac{\partial g_{kt}}{\partial x^l} = \frac{1}{2|g|} \frac{\partial |g|}{\partial x^l} = \frac{1}{2|g|} |g|_{,l} \\ &= \frac{1}{2} (\log |g|)_{,l} = (\log \sqrt{|g|})_{,l} = \frac{(\sqrt{|g|})_{,l}}{\sqrt{|g|}}. \end{aligned} \quad (6.28)$$

Note that if the determinant g is negative, as it generally is in relativity theory, the $\sqrt{|g|}$ in (6.18) must be replaced by $\sqrt{-|g|}$; equivalently we may interpret $\sqrt{|g|}$ as being the root of the absolute value of the determinant.

These expressions are often of use. With the final form in (6.28) we may return to the expression (6.25) for the divergence and write it as

$$\begin{aligned} B^k{}_{;k} &= B^k{}_{,k} + \Gamma^k_{kl} B^l = B^k{}_{,k} + \frac{(\sqrt{|g|})_{,k}}{\sqrt{|g|}} B^k \\ &= \frac{1}{\sqrt{|g|}} \left(\sqrt{|g|} B^j \right)_{,j} \quad \text{generalized divergence.} \end{aligned} \quad (6.29)$$

Thus we have expressed the divergence in an elegant form that contains no connection but only the metric determinant and an ordinary derivative; one need not calculate the connections.

The form for the divergence (6.29) and the invariant volume element discussed in Sect. 4.7 combine beautifully in giving a covariant version of Gauss's law for integrals; see Exercise 6.9.

In elementary vector calculus the Laplacian is defined as the divergence of the gradient of a scalar,

$$\operatorname{div} \operatorname{grad} \phi = \nabla \cdot \nabla \phi = \nabla^2 \phi = \frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z^2} \quad \text{Laplacian.} \quad (6.30)$$

The natural generalization of this is to use the above definition of divergence on the gradient of a scalar, or

$$\nabla^2 \phi = (g^{ij} \phi_{,j})_{,i}, \quad \text{generalized Laplacian.} \quad (6.31)$$

As we have shown for the divergence this may be written without connections as

$$(g^{ij} \phi_{,j})_{,i} = \frac{1}{\sqrt{|g|}} \left(\sqrt{|g|} g^{ij} \phi_{,j} \right)_{,i}. \quad (6.32)$$

This form is quite useful for doing vector analysis in a curvilinear coordinate system, and gives the familiar textbook expressions for the Laplacian in spherical and cylindrical coordinates with ease. It is important in tensor analysis in general relativity.

Example 6.1 Let us work out the simple but nontrivial example of the Laplacian in polar coordinates. Call the scalar function f . From Example 4.1 we have the metric and its inverse,

$$g_{ij} = \begin{pmatrix} 1 & 0 \\ 0 & \rho^2 \end{pmatrix}, \quad g^{ij} = \begin{pmatrix} 1 & 0 \\ 0 & 1/\rho^2 \end{pmatrix}, \quad \sqrt{|g|} = \rho. \quad (6.33)$$

The covariant gradient and the corresponding contravariant vector are

$$f_{,k} = (f_{,\rho}, f_{,\varphi}), \quad g^{ik} f_{,k} = (f_{,\rho}, f_{,\varphi}/\rho^2). \quad (6.34)$$

Substituting these into (6.32), we find for the Laplacian the well-known expression,

$$\nabla^2 f = \frac{1}{\rho} (\rho f_{,\rho})_{,\rho} + \frac{1}{\rho^2} f_{,\varphi,\varphi} = f_{,\rho,\rho} + \frac{1}{\rho} f_{,\rho} + \frac{1}{\rho^2} f_{,\varphi,\varphi}. \quad (6.35)$$

Appendix 1: Curve Derivatives as Vectors

There is a somewhat more sophisticated notation that the reader may encounter concerning the abstract approach to vectors in Sect. 6.2. We will only mention here the basic concept and the nomenclature (Misner 1973; Ohanian 1994).

Consider a curve C parameterized by its arc length or other invariant parameter λ as in Fig. 6.2. We could define a function $f(\lambda)$ along the curve and thereby its derivative. We do not even need coordinates to think about this construction. For example the space could be a 2-surface and we could mark C on it with a pen, then measure λ along it with a flexible tape. We define the tangent vector $\vec{t}$ to the curve at P as the directional derivative operator on any such function with respect to λ

$$\vec{t} = \frac{d}{d\lambda}, \quad \vec{t}(f) = \frac{df}{d\lambda}. \quad (6.36)$$

This definition, as we will see, implies that the components of $\vec{t}$ are the same objects that we have been using as the components of a tangent vector; to see this we express the curve derivative using the chain rule and compare the tangent vector expressions with what we have used previously, as in (6.18),

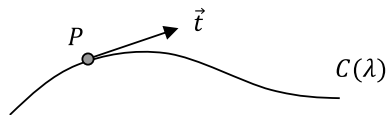


Fig. 6.2 The curve C has a tangent vector $\vec{t}$ at the point P . It need not be defined in terms of a coordinate system, but it can be if desired

$$\vec{t} = \frac{d}{d\lambda} = \frac{dx^i}{d\lambda} \frac{\partial}{\partial x^i} \quad \text{versus} \quad \vec{t} = \frac{dx^i}{ds} \vec{e}_i, \quad \text{so} \quad \frac{\partial}{\partial x^i} \leftrightarrow \vec{e}_i. \quad (6.37)$$

That is the curve derivative operators along coordinate lines act just like the coordinate basis vectors, and the curve derivative itself acts like a vector.

Appendix 2: p -Forms and Exterior Derivatives

The reader may encounter objects called p -forms in the literature. Here we will only mention the basic concept and the nomenclature (Misner 1973; Ohanian 1994).

The 1-forms we have used naturally generalize to these larger p -forms, which are useful in some physics applications. A 2-form is defined as the antisymmetrized exterior product of 1-forms, say $\tilde{\alpha}$ and $\tilde{\beta}$. Specifically

$$\tilde{\alpha} \wedge \tilde{\beta} \equiv \tilde{\alpha} \otimes \tilde{\beta} - \tilde{\beta} \otimes \tilde{\alpha}. \quad (6.38)$$

If the 1-forms are expanded in terms of a coordinate basis then the 2-form may be written

$$\tilde{\alpha} = \alpha_\mu \tilde{dx}^\mu, \quad \tilde{\beta} = \beta_\sigma \tilde{dx}^\sigma, \quad \tilde{\alpha} \wedge \tilde{\beta} = \frac{1}{2} (\alpha_\mu \beta_\sigma - \beta_\mu \alpha_\sigma) (\tilde{dx}^\mu \wedge \tilde{dx}^\sigma). \quad (6.39)$$

The general p -form is defined as the anti-symmetric product of p such factors,

$$\tilde{\alpha} \equiv \frac{1}{p!} (\alpha_{\mu\sigma\dots\lambda}) (\tilde{dx}^\mu \wedge \tilde{dx}^\sigma \dots \wedge \tilde{dx}^\lambda), \quad (6.40)$$

where the coefficient set is anti-symmetric.

The exterior derivative is defined as the only $p + 1$ form that can be built from a p -form by differentiation; it is explicitly

$$d\tilde{\alpha} \equiv \frac{1}{p!} (\alpha_{\mu\sigma\dots\lambda,\beta}) (\tilde{dx}^\beta \wedge \tilde{dx}^\mu \wedge \tilde{dx}^\sigma \dots \wedge \tilde{dx}^\lambda). \quad (6.41)$$

If we differentiate the exterior derivative again we obviously get zero because the coefficient set becomes symmetric in two indices, and the basis form is anti-symmetric. We may abbreviate this statement as $d(d\tilde{\alpha}) = 0$.

Such p -forms are useful in physics where antisymmetric tensors occur; in electromagnetism they allow an elegant treatment of Maxwell's equations. A few brief comments on p -forms in four dimensions are thus in order: A 0-form is simply defined as a scalar and has 1 independent component. The 1-forms we have been using have four independent components. The 2-forms have the same number of components as an anti-symmetric 4 by 4 matrix, which is 6; this is the number of components in the electromagnetic Maxwell tensor. The 3-forms can be labeled by the missing

index and thus have 4 components. A 4-form coefficient array must be a multiple of the Levi-Cevita epsilon, which is discussed in Exercise 6.6, so the 4-form has only 1 independent component.

We will develop gravitational theory in this book without the use of p -forms so we will not discuss them further. For the reader interested in the gravitational field associated with the electromagnetic field they can be useful (Adler 1975; Misner 1973).

Exercises

- 6.1 Let us study a vector field in Euclidean 2-space. In Cartesian coordinates take the field to be $V^i(x^j) = (1, 1)$. This is a constant field represented by arrows at 45° throughout space.
 - (a) Transform the vector field to polar coordinates and call it $\bar{V}^i$. (We obtained the transformation matrix in Example 4.1.) Note that it does not have constant components. Lower the index and form the covariant field $\bar{V}_k$, which also does not have constant components.
 - (b) Sketch the field in terms of arrows in polar coordinates, and see that the same picture results as with Cartesian coordinates. (Use the ideas discussed in Example 4.4.)
 - (c) Calculate the covariant derivative $V^i{}_{;k}$ in Cartesian coordinates, which of course is trivial. What is it in polar coordinates.
- 6.2 Consider the vector field $V^i(x^j) = (1, 0)$ in polar coordinates. Draw a picture of it. Calculate its covariant derivative and its divergence.
- 6.3 For the covariant Laplacian in (6.32) ...
 - (a) Calculate the Laplacian for a coordinate system in which the metric is constant.
 - (b) Calculate it for Euclidean 3-space with cylindrical coordinates.
 - (c) Calculate it for Euclidean 3-space with spherical coordinates.
- 6.4 Calculate the Laplacian specifically for Minkowski space using Cartesian coordinates. This is also called the d'Alembertian operator. Setting the d'Alembertian of a function f equal to zero gives the scalar wave equation. Show that one important type of solution is $f(x - ct)$, where f is any twice differentiable function. We will discuss this solution at length when we study gravitational waves in Chap. 11.
- 6.5 In Exercise 5.3 we studied a line element of the form $ds^2 = f(\rho)^2 d\rho^2 + \rho^2 d\varphi^2$.
 - (a) What is the square root of the metric determinant, and what is the invariant volume element?
 - (b) Consider the special case $f(\rho) = 1 - a/\rho$, where a is a positive constant. Note that the 1,1 metric component can be zero. What do you suppose this means? Hint: calculate $\sqrt{|g|}$ and think about the invariant volume element. What peculiar features does this imply for the line element? A similar peculiarity occurs for black holes, which we will discuss in Part III.

- (c) Refer back to the discussion in Example 4.2 and especially (4.10). Can this $f(\rho)$ be the metric of a 2-surface imbedded in Euclidean 3-space?
- 6.6 The Levi-Cevita epsilon occurs often in matrix and tensor theory; in n dimensions it has n indices and is defined in terms of its indices as

$$\epsilon_{\alpha\beta\gamma\dots\tau} = \begin{cases} 0 & \text{if 2 indices are equal} \\ \pm 1 & \text{for even/odd permutations of } 1 \ 2 \ \dots \ n \end{cases}$$

The epsilon is not a tensor, but the object quantity $e_{\alpha\beta\gamma\dots\tau} = \sqrt{|g|}\epsilon_{\alpha\beta\gamma\dots\tau}$ is a tensor. Show this. Hint: Express the determinant of a matrix using the epsilon.

- 6.7 In three dimensions show that the epsilon obeys

$$\epsilon_{ijk}\epsilon_{imn} = \delta_{jm}\delta_{kn} - \delta_{jn}\delta_{km}, \quad \text{sum over } i.$$

- 6.8 Show that the covariant derivative of a coordinate basis vector may be written as

$$\nabla \vec{e}_n = \Gamma_{nk}^i \vec{e}_i \otimes \tilde{d}x^k.$$

- 6.9 Let us see how nicely the divergence and the invariant volume element fit together. To do this consider in Euclidean 3-space the volume integral of the divergence of a vector. Use the divergence expression (6.29) and the invariant volume element expression (4.81) and see how the volume integral becomes a surface integral, that is Gauss's Theorem. Is it clear how useful this is for a spherically symmetric system in spherical coordinates?
- 6.10 For some familiar important cases let's solve Laplace's equation, that is setting the Laplacian of a scalar function equal to zero. Using the results of Exercise 6.3...
- (a) Solve it for a spherically symmetric system in spherical coordinates. Notice that you must allow a singularity or the only solution is zero! This solution is useful in Part III.
- (b) Solve it for a cylindrically symmetric system in cylindrical coordinates.

Part III

General Relativity

We have now studied enough vector and tensor analysis that we may apply the mathematics to physics. We begin by looking at the familiar classical gravitational force from a new perspective, as a geometric effect. To develop this idea fully, we return briefly to mathematics and study curvature in a Riemann space, from which the general relativistic field equations of gravity follow in a natural way.

As the most fundamental application of the field equations, we then study the spherically symmetric gravitational field solution of Schwarzschild, which describes the solar system quite well. This is the oldest and most important exact solution in the theory. It provides a description of the solar system that has been tested to impressive accuracy.

Then we progress to much stronger gravitational fields, such as those of a neutron star or a black hole, that is a collapsed star. To study the collapse of matter to a black hole we consider the classic example of a dust ball with negligible pressure.

Next we consider black holes themselves and some of their extraordinary properties. One of the most interesting properties that we study is their emission of radiation like a black body, the Hawking radiation.

Finally we consider weak gravitational fields, for which the theory becomes linear. As an important application we study gravitational waves; these have been detected and a new window on the universe has thereby been opened. In particular the waves from the merger of black holes and neutron stars have been detected so the fields of gravitational wave physics, black hole physics and neutron star physics have expanded and become closely connected.

Chapter 7

Classical Gravity and Geometry



Abstract In this chapter we look at the familiar classical gravitational force from a novel perspective, as a geometric effect. This perspective is motivated by the equivalence principle, the close similarity of gravitational effects to the effects of acceleration. As an application of the geometric view the gravitational redshift can be easily derived.

7.1 Newtonian Gravity

Classical or Newtonian gravitational theory is well-known to almost all physicists, so only a short review need be given here. For more detail see Chap. 1 of Ohanian (1994) and Chap. 12 of Misner (1973). Our review is focused on the troubles with the theory. The basic postulate is the inverse square law of Newton, in which the force of attraction between point masses M and m separated by distance r is given by

$$\vec{F} = -\frac{GMm}{r^2}\hat{r}, \quad G = 6.672 \times 10^{-11} \text{ N m}^2/\text{kg}, \quad \text{Newtonian gravity.} \quad (7.1)$$

Notice how similar this is to the Coulomb force law of electrostatics for charges q and Q ,

$$\vec{F} = \frac{Qq}{4\pi\epsilon_0 r^2}\hat{r}, \quad \frac{1}{4\pi\epsilon_0} = 8.99 \times 10^9 \text{ N m}^2/\text{C}^2, \quad \text{electrostatics.} \quad (7.2)$$

The main difference is that in electrostatics the charges Q and q can have either sign and the force may thus be attractive or repulsive. One may thus develop classical gravitational theory in close analogy with electrostatics, using the correspondence *mass* $\leftrightarrow$ *charge* and $G \leftrightarrow 1/4\pi\epsilon_0$. Only the signs require some care. For example, we define a gravitational vector field $\vec{g}$ by $\vec{F} = m\vec{g}$, so for a point mass

$$\vec{g} = -\frac{GM}{r^2}\hat{r}. \quad (7.3)$$

A gravitational *potential* ϕ is defined by $\vec{g} = -\nabla\phi$ and a gravitational *potential energy* by $V = m\phi$, so for a point mass

$$\phi = -\frac{GM}{r}, \quad V = -\frac{GMm}{r}. \quad (7.4)$$

For a continuous distribution of matter we may superpose point masses with a mass density function ρ and find

$$\phi(\vec{r}) = -G \int \frac{\rho(\vec{r}')}{|\vec{r} - \vec{r}'|} d^3r'. \quad (7.5)$$

The last expression may be used to obtain Poisson's equation for the potential,

$$\nabla^2\phi = 4\pi G\rho. \quad (7.6)$$

Alternatively, we may postulate Poisson's equation and obtain the force laws, just as in electrostatics, and develop the whole theory on that basis.

Newtonian gravitational theory is extraordinarily accurate, and for over 200 years was used to study the solar system with no known errors. Despite the success in predicting empirical observations there are two defects with the theory which led to its abandonment and the adoption of Einstein's relativistic theory of gravity. These are:

- (1) *Classical gravity is instantaneous*: the distance in (7.1) is the relative separation of the masses when the mass m feels the force exerted by M . But special relativity is not compatible with such action-at-a-distance or instantaneous propagation, as we will discuss. Of course, this was only seen to be a defect after special relativity was developed in 1905.
- (2) *The masses in (7.1) are the same as the inertial masses*: these are defined in terms of resistance to acceleration via Newton's second law, $\vec{F} = m\vec{a}$. Why should the same inertial mass produce a gravitational field? The analogy with electrostatics is useful here; the charge of a particle which produces the electric field is independent of the inertial mass of the particle, so why should the "gravitational mass" of the particle which produces the gravitational field be the same as the inertial mass of the particle?

Notice that defect (1) involves a measurably real physics problem, while (2) only involves a conceptual quandry, that an important and fundamental equality is not explained by the theory but merely postulated.

Let us look at defect (1) a little further in the light of special relativity. We may set up a thought experiment or gedanken experiment, as Einstein was fond of doing. Suppose we are at the position of m , and a colleague wiggles the mass M ; according to (7.1) we would see the effect immediately, so the force propagates at infinite velocity over the distance $r = \Delta x$ in time $\Delta t = 0$ as in Fig. 7.1.

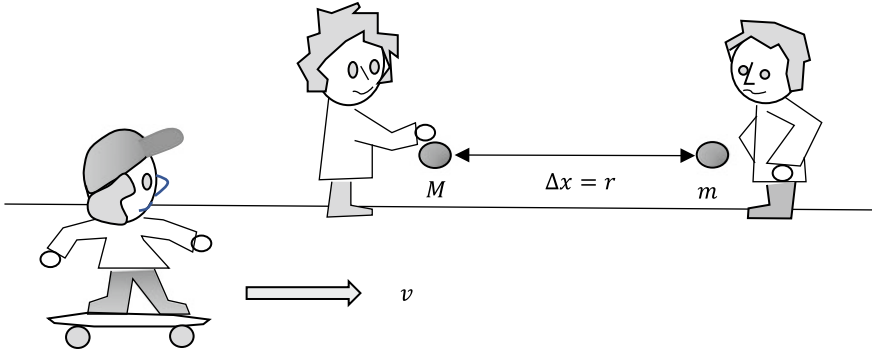


Fig. 7.1 Point masses attract each other by the inverse square law. The moving observer will see a very peculiar effect, as discussed below

Another observer in a system moving past us at velocity v would not see the same space and time intervals however, and by the Lorentz transformation (Sect. 1.2) he would instead see

$$\begin{aligned} c\Delta t' &= \gamma c\Delta t - \beta\gamma\Delta x = -\beta\gamma\Delta x, \\ \Delta x' &= -\beta\gamma c\Delta t + \gamma\Delta x = \gamma\Delta x. \end{aligned} \quad (7.7)$$

Here, as in Chap. 1 the definitions are $\beta = v/c$ and $\gamma = 1/\sqrt{1 - \beta^2}$. That is, the moving observer would see a *negative* time difference: for him the signal would reach us before being sent by our colleague. We refer to this as a violation of causality, and the situation is so peculiar that it is generally considered unacceptable. Thus no “action at a distance” type theory is acceptable since it cannot be consistent with relativity.

Let us return to (7.7) and see how fast the signal may propagate so as not to reverse the sign of the time interval and violate causality. That is we demand that the time interval seen by the moving observer according to (7.7) be positive, so that

$$c\Delta t' = \gamma c\Delta t - \beta\gamma\Delta x \geq 0, \quad \text{thus} \quad \beta \frac{\Delta x}{\Delta t} = \beta v_{\text{prop}} \leq c. \quad (7.8)$$

But β is whatever velocity the moving observer may have, which is any value up to 1. Thus

$$v_{\text{prop}} \leq c. \quad (7.9)$$

That is the propagation velocity cannot exceed the speed of light, just as moving observers and objects may not exceed it. In order to make gravity consistent with

special relativity the fundamental equation (7.1) must be modified so that gravitational effects propagate at c or less. The situation is rather remarkable: the theory must be changed despite a lack of any experimental evidence that it is wrong.

Problem (2), the equality of the inertial and gravitational masses, is curious in that it leads to fundamentally strange consequences, which are quite well-known and familiar. From the above definitions we may write Newton's second law for the acceleration of a test body in a gravitational field

$$\vec{F} = -m\nabla\phi = m\vec{a}, \quad \text{so} \quad \frac{d^2x^j}{dt^2} = -\phi_{,j}. \quad (7.10)$$

Because of the equality of inertial and gravitational mass the mass of the test body cancels from the equation for the acceleration and the acceleration is independent of it. Thus, for example, if two objects of different mass in the earth's field begin at the same position with the same velocity they will follow the same trajectory. Similarly, the paths of planets around the sun are independent of the planet mass. Astronauts inside a spacecraft in orbit follow the same trajectory as the spacecraft and therefore float freely inside the craft. We call this free-fall.

The fact that different bodies fall at the same rate in a gravitational field is often referred to as the *universality of free fall* or the *weak equivalence principle*. We will say more about it in the following section when we further discuss the equivalence principle and when we study the intrinsic signature of gravity in general relativity in Sect. 8.5.

It is worth emphasizing that the equality of inertial and gravitational mass is subject to experimental test of very high accuracy. Eotvos in the early twentieth century showed that the two are equal within about a part in 10^8 , while more recently Dicke et al. have increased the accuracy to better than a part in about 10^{12} (Eotvos 1922; Will 2014). There are presently proposals to test the equality in a spacecraft with an accuracy of about a part in 10^{17} (Will 2014).

In the context of Newtonian theory the question of why such an extraordinary situation should occur is a deep mystery. By contrast it follows easily and naturally from a geometrical viewpoint, and was one of the main guides used by Einstein in developing the relativistic theory of gravity (Zee 1989).

7.2 The Equivalence Principle

Let us follow Einstein in his reasoning concerning the equivalence principle (EP) using gedanken experiments. We begin by putting one observer in a lab on the earth and one in an identical lab in a rocket ship in space accelerating at g , as shown in Fig. 7.2. (Einstein used an elevator rather than a rocket.)

In the two labs we then do various mechanics experiments, like dropping balls, weighing objects, setting up levers and inclined plane systems etc. In the earth lab a ball accelerates downward due to the force of gravity, and independent of its mass. In

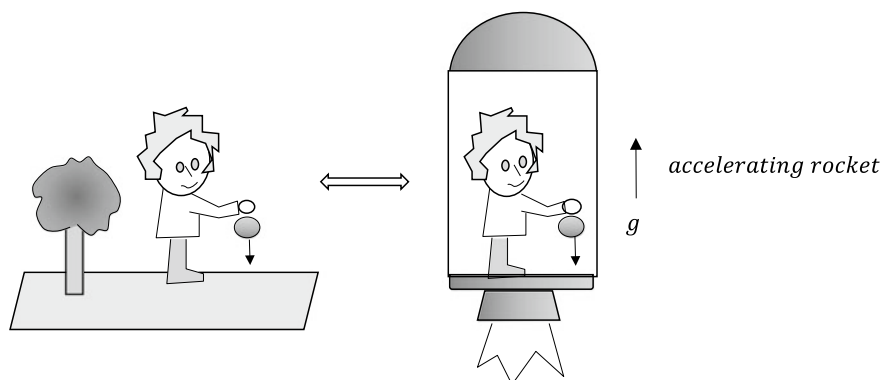


Fig. 7.2 In the earth lab and the accelerated lab observers see the same mechanical phenomena

the accelerated lab a ball moves at constant velocity and the floor accelerates upward to catch it, clearly independent of its mass. A moment's thought assures us that we cannot tell by such experiments if we are in the earth or the space lab—there is an *equivalence* between phenomena in an *accelerated* system and in a *gravitational field*. This is the equivalence principle in its simplest form. Note carefully that it obviously can hold only in a very *small* lab over which the gravitational field may be considered *uniform*: it is a local principle. We will discuss this further below.

There is an obvious converse to the above equivalence. Let the earth observer fall freely, and turn off the engine of the rocket ship. This scenario is shown in Fig. 7.3.

Then one again sees the same phenomena in the two labs: things float about freely. We may view this as “turning off” the gravitational field in the earth lab.

Einstein considered this to be a great insight, that the gravitational field in a small region of spacetime is equivalent to acceleration of the lab system, and may be turned off by a different choice of the lab system. It is thus very like a fictitious force which one often encounters in classical physics, such as centrifugal and Coriolis forces;

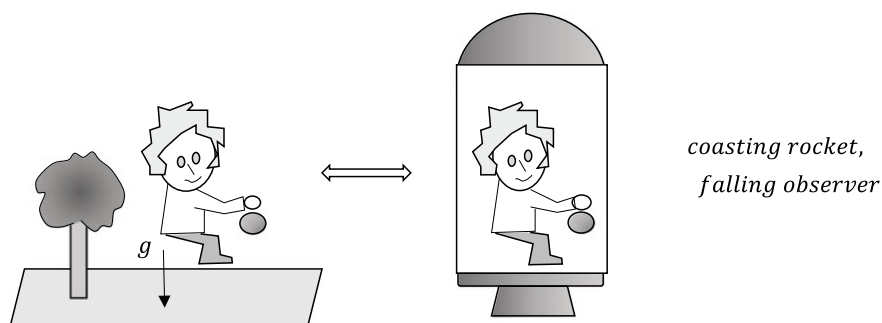


Fig. 7.3 The converse of Fig. 7.2, in which the observer falls freely and the rocket engine is turned off. Everything floats freely

such forces may also be turned off by going to a different lab or coordinate system; moreover, and very importantly, such fictitious forces are represented by connection terms in the classical equations of motion (5.65). Since gravity is similar to fictitious forces in that it is proportional to mass, and can be transformed away, might it then be represented by connection terms in equations of motion and thereby be thought of as a geometric effect? The answer is of course “yes” as we will show in the next section.

An important caveat is associated with the equivalence principle. We again emphasize that the lab must be considered so small that the gravitational field is effectively uniform over it. In a larger lab there is an obvious difference between the earth lab and the accelerated lab: two balls falling in the earth lab will converge ever so slightly as they fall toward the center of the earth, and in the rocket lab they will not (see Fig. 7.4). The slight convergence is due to the fact that the earth lab has a gravitational field with a gradient and consequent *tidal* forces. These tidal forces are the intrinsic signature of the gravitational field, not the acceleration of a test body. Indeed, this fact is crucially important; in relativistic gravity we will see that the Riemann curvature tensor is the analog of Newtonian tidal forces and is the signature of the gravitational field or spacetime curvature. We will study and make further use of this fact in Chap. 8.

Einstein elevated the principle of equivalence from an observation about mechanics to a general principle of physics; he assumed that not only mechanical effects like those we mentioned above but *all physical effects* will be the same in a gravitational field as in the equivalent accelerating system (Kenyon 1990; Will 1993). This is often called the Einstein equivalence principle. For example, one consequence is that light must be deflected in a gravitational field, because in the equivalent accelerating lab a beam of light waves sent across the lab will clearly be seen to curve downward.

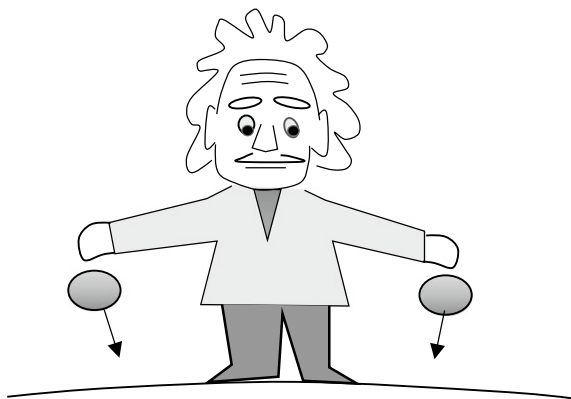


Fig. 7.4 The equivalence principle does not apply if the lab is large enough that nonuniformity in the gravitational field is detectable. The balls are seen to converge toward the center of the earth

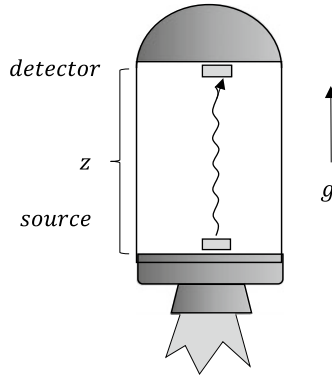


Fig. 7.5 The Doppler shift or redshift experimental layout in the accelerated lab

Example 7.1 We analyze the gravitational redshift using the equivalence principle.

Consider on the floor of an accelerated lab a light source which emits a pulse of light of wavelength λ , as shown in Fig. 7.5. This moves upward a distance z to be by a detector in about time $\Delta t = z/c$. During this time the lab has accelerated and the detector is moving upward slightly faster than the source was moving when the light was emitted, with $\Delta v = g \Delta t$. Thus there will be a Doppler shift in the light, given by (1.27), of approximately

$$\lambda_{\text{obs}} - \lambda = \Delta\lambda = \frac{\Delta v}{c} \lambda, \quad \frac{\Delta\lambda}{\lambda} = \frac{\Delta v}{c} = \frac{g \Delta t}{c} = \frac{gz}{c^2}. \quad (7.11)$$

The equivalence principle tells us that we see the same effect in the earth lab, the light redshifted to longer wavelength. Moreover in the earth-based lab $\Delta\phi = gz$ is the difference between the gravitational potential at the source and the receiver, so we may write

$$\frac{\Delta\lambda}{\lambda} = \frac{\Delta\phi}{c^2}. \quad (7.12)$$

This prediction has been tested experimentally. A very accurate test used a microwave system in a high-altitude rocket sent to about 10^4 km, which gave a result in agreement with the above prediction to better than about a part in 10^4 (Vessot 1980).

It is worth noting that any redshift experiment only tests the equivalence principle and the general ideas of general relativity, but since we have not yet presented the field equations it is clearly not a test of the field equations and the full relativity theory.

There has been a great deal of both theoretical and experimental work done on the equivalence principle and a variety of versions have been discussed, most notably the weak equivalence principle and the Einstein equivalence principle that we discussed above. We will use here only the most basic, the universality of free fall, or weak equivalence principle. For a more detailed discussion of the various statements of the principle and relevant experiments on this important topic see Will (1993, 2014).

7.3 Gravity as a Geometric Phenomenon

As sketched above Newtonian gravitational theory is based on distances in 3-dimensional space and an absolute universal time. It is therefore quite remarkable that the conceptual framework of Chap. 5 for affine and metric spaces combined with some basic ideas of special relativity leads to classical gravity as an approximation for slow motion in weak fields.

Let us first note how the concept of vector parallel displacement can be naturally related to the concept of a classical force. Apply the basic vector displacement expression (5.5) to the 4-vector velocity $u^\beta = dx^\beta/d\tau$ of a body in the spacetime of special relativity, and consider vector transplantation in the time direction by cdt for low velocity $u^i \ll u^0$ and $u^0 \cong c$. This gives for the approximate change in a space component of the velocity,

$$du^j \cong -\Gamma_{00}^j c^2 dt, \quad \frac{du^j}{dt} = \frac{d^2 x^j}{dt^2} = a^j \cong -\Gamma_{00}^j c^2. \quad (7.13)$$

This is just Newton's second law $\vec{F} = m\vec{a}$, where the affine connection Γ_{00}^j plays the role of a force per unit mass. Notice that it is important that the transplantation is done in the 4-dimensional spacetime of relativity rather than the 3-space of classical physics, and also note that the mass of the body does not explicitly appear so the EP is implied.

Let us pursue this geometric viewpoint further, but more precisely and explicitly. We saw in the appendix on classical mechanics in Chap. 5 that fictitious forces are represented by connections in the equations of motion (5.65). Recall that such forces are called fictitious because they are proportional to the mass of the test body and may be transformed away by a different choice of lab system or coordinates. But we have just seen that the force of gravity also is proportional to the mass of the test body and may be transformed away by a different choice of lab system. It is natural that we then try to represent gravity as a fictitious force as in (5.65) (Adler 1975; Zee 1989).

Here are the four rules for this analysis:

- (1) We use the Lorentz metric of special relativity, but modify it a small amount

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}, \quad h_{\mu\nu} \ll 1. \quad (7.14)$$

The $h_{\mu\nu}$ represents a weak gravitational field.

- (2) We take $h_{\mu\nu}$ to be time independent or very slowly varying, and also diagonal as we will justify in a later section (see also Exercise 7.9).
- (3) The 3-velocity of all bodies considered is small, that is $\beta \ll 1$.
- (4) We assume the equation of motion for a body is the geodesic equation because a geodesic is the only privileged curve in a metric space.

In the geodesic equation

$$\ddot{x}^\mu + \Gamma_{\alpha\beta}^\mu \dot{x}^\alpha \dot{x}^\beta = 0 \quad (7.15)$$

the dot signifies a derivative with respect to the line element, whereas the classical theory involves time derivatives. We can relate the two using the line element as we did in Part I on special relativity. From (7.14) the line element along the geodesic is

$$\begin{aligned} ds^2 &= c^2 dt^2 - (d\vec{x})^2 + h_{\mu\nu} dx^\mu dx^\nu \\ &= (1 - \beta^2 + h_{00})c^2 dt^2 = (1 + \varepsilon)^2 c^2 dt^2, \\ 1 + \varepsilon &\equiv \sqrt{1 + h_{00} - \beta^2} \cong 1 + \frac{h_{00}}{2} - \frac{\beta^2}{2}, \end{aligned} \quad (7.16)$$

where β is the velocity over c along the geodesic. We have retained second order terms in the velocity and first order terms in $h_{\mu\nu}$; for bodies in the solar system the dimensionless quantities β^2 and h_{00} are comparable and very small, as we will later discuss (see Exercise 7.5). From (7.16) we find the relation between the proper time and coordinate time derivatives to be approximately

$$\frac{ds}{dt} = (1 + \varepsilon)c, \quad \frac{d}{ds} = \frac{dt}{ds} \frac{d}{dt} = \frac{1}{1 + \varepsilon} \frac{1}{c} \frac{d}{dt} = (1 - \varepsilon) \frac{1}{c} \frac{d}{dt}. \quad (7.17)$$

Using this relation we find, to lowest order in β^2 and h_{00} and ε ,

$$\Gamma_{\alpha\beta}^i \dot{x}^\alpha \dot{x}^\beta = (1 - 2\varepsilon) \left[\Gamma_{00}^i + 2\Gamma_{0j}^i \frac{v^j}{c} \right]. \quad (7.18)$$

Similarly we find with a little algebra,

$$\ddot{x}^i = (1 - 2\varepsilon) \frac{1}{c^2} \frac{d^2 x^i}{dt^2}, \quad (7.19)$$

where we have used the assumption (2), that h_{00} is independent of time. Combining (7.18) and (7.19) we obtain the approximate geodesic equation in terms of time derivatives,

$$\frac{d^2 x^i}{dt^2} + c^2 \left[\Gamma_{00}^i + 2\Gamma_{0j}^i \frac{v^j}{c} \right] = 0. \quad (7.20)$$

This is a more accurate and justified version of (7.13). It is worth pondering for a moment. It says that within the approximation framework that we have set up the gravitational force is represented by connections, analogous to the fictitious forces of classical mechanics. To make this correspondence it was necessary to use the 4-dimensions of spacetime in special relativity. Equation (7.20) clearly shows how the ideas of geometry and classical forces and accelerations are related, and that the motion is independent of the mass of the body.

To finish our task and relate the geometric view to the classical potential we need only evaluate the connections in (7.20). We find from their definition

$$\begin{aligned}\Gamma_{00}^i &= \frac{1}{2}\eta^{ik}(h_{0k,0} + h_{k0,0} - h_{00,k}) = \frac{1}{2}h_{00,i}, \\ \Gamma_{0j}^i &= \frac{1}{2}\eta^{ik}(h_{0k,j} + h_{kj,0} - h_{0j,k}) = 0,\end{aligned}\quad (7.21)$$

where we have used the time independence and the diagonal nature of the metric (see Exercises 7.8 and 7.9). We finally bring everything together and substitute (7.21) into (7.20) to obtain

$$\frac{d^2x^i}{dt^2} = -\frac{1}{2}c^2h_{00,i}. \quad (7.22)$$

This is a wonderful result. It is identical to the classical equation (7.10) if we identify

$$\phi_{,i} = \frac{1}{2}c^2h_{00,i}, \quad \text{so that } g_{00} = 1 + h_{00} = 1 + \frac{2\phi}{c^2}. \quad (7.23)$$

Therefore, in summary, we get classical gravitational theory as the weak field and low velocity limit of a geometric theory provided that the g_{00} component of the metric is related to the classical potential by (7.23). We emphasize that it is the *time* part of the metric that is important and the other components of the metric play a lesser role in this correspondence. See Exercises 7.8 and 7.9 for further comments on an analysis to higher order.

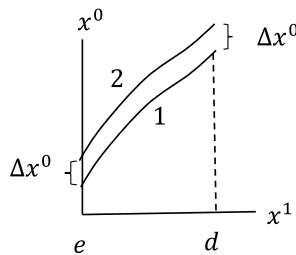


Fig. 7.6 An emitting atom at e sends radiation to a detector at d . The trajectories of the rays are simply shifted in time

Example 7.2 Let us return to the gravitational redshift. We have already estimated the redshift using the equivalence principle, but with the above result relating the metric to the gravitational potential we may derive it in a more precise and general geometric way. Consider a stationary emitter of radiation, such as an atom, at position e and a detector at d , as shown in Fig. 7.6.

Suppose the beginning of a cycle number 1 leaves the emitter at coordinate time x^0 and the beginning of another cycle number 2 leaves a very short coordinate time Δx^0 later. The paths of these travel through 3-space as a function of time; whatever determines the path of number 1 it is clear that number 2 will encounter very nearly the same conditions since it left a very short time later, so 1 and 2 will follow the same path but with number 2 displaced uniformly upwards by Δx^0 , as shown in the figure. Thus the *coordinate* time period of the radiation will be the same at e and d . But the coordinates are merely markers or labels for points in spacetime and have no direct physical meaning. As in special relativity the *proper* time $\Delta\tau = \Delta s/c$ is what has physical meaning. The relations between proper and coordinate time at the stationary emitter and detector are

$$c\Delta\tau_e = \sqrt{g_{00}(e)}\Delta x^0, \quad c\Delta\tau_d = \sqrt{g_{00}(d)}\Delta x^0. \quad (7.24)$$

Thus we obtain a relation between the period of the radiation at the emitter and at the detector,

$$\frac{\Delta\tau_d}{\Delta\tau_e} = \frac{\sqrt{g_{00}(d)}}{\sqrt{g_{00}(e)}}. \quad (7.25)$$

This is quite general and holds for widely separated emitter and detector, unlike the equivalence principle derivation. You should think about the implication of (7.25) when g_{00} at the emitter or detector is very small or zero.

To show that this is consistent with the equivalence principle result (7.12) we use the relation between g_{00} and the gravitational potential in (7.23) and expand, assuming a weak field, to get

$$\Delta\tau_d = \frac{\sqrt{g_{00}(d)}}{\sqrt{g_{00}(e)}}\Delta\tau_e = \frac{\sqrt{1 + 2\phi(d)/c^2}}{\sqrt{1 + 2\phi(e)/c^2}}\Delta\tau_e = \left(1 + \frac{\Delta\phi}{c^2}\right)\Delta\tau_e, \quad (7.26)$$

or in terms of the wavelength

$$\frac{\Delta\tau_d - \Delta\tau_e}{\Delta\tau_e} = \frac{\Delta\lambda}{\lambda} = \frac{\Delta\phi}{c^2}. \quad (7.27)$$

Thus the results in (7.12) and (7.25) are consistent.

It is worth pondering for a moment the conceptual view of gravity that is provided by the above results. For a weak gravitational field bodies follow geodesics in the spacetime of special relativity with a small correction: the metric is modified so that their internal clocks tick at a slightly different rate depending on their proximity to matter, with time intervals $d\tau = ds/c$ determined from (7.23) as

$$d\tau = \left(1 + \frac{\phi}{c^2}\right) dt. \quad (7.28)$$

The potential ϕ is taken to be zero far from all sources of gravity.

Exercises

- 7.1 Using Poisson's equation of classical gravitational theory (7.6) calculate the potential ϕ and the field $\vec{g}$ for a space filled with constant density matter. Assume that the field is spherically symmetric about some arbitrary origin. Notice that the uniform distribution of the matter appears to have a greater degree of symmetry than the gravitational field. Does this bother you?
- 7.2 A common theme in science fiction is negative matter which falls upwards in a gravitational field. How much general relativity do you need to know in order to be very dubious of such a notion?
- 7.3 What is the gravitational redshift between a point on the earth's surface and a point on the sun's surface? What is it between two points separated by a vertical 100 m on the surface of the earth? What is it between the earth's surface and a point at 10,000 km altitude? See Vessot (1980).
- 7.4 Does an experimental test of the redshift really test general relativity theory? What if the measurement is extremely accurate? What would happen if g_{00} were zero at the point of emission? We will discuss just this situation in Chap. 10.
- 7.5 In our low velocity and weak field discussion the combination of velocity squared and field strength that appears in (7.16) is $h_{00} - \beta^2$. Show that h_{00} and β^2 are related and comparable for planets in circular orbit around the sun.
- 7.6 When we studied the Newtonian limit of (7.15) we only considered the space parts, with $\mu = i$. Show that the time equation, $\mu = 0$, is consistent but does not give us any interesting new information.
- 7.7 We obtained the gravitational redshift formula using two different methods. Add a third by considering a photon moving upward in the field of the earth and losing energy as it rises. You can do this heuristically by assigning the Planck energy $E = h\nu$ to the photon, with a corresponding effective mass $m_{\text{eff}} = E/c^2$.
- 7.8 In obtaining the equation of motion (7.22) we assumed that the metric was diagonal. Repeat the derivation without this assumption; specifically, allow the h_{0j} to be nonzero so that the second equation in (7.21) no longer holds and a velocity dependent force is added to (7.22).
- 7.9 Continue studying the velocity dependent force of Exercise 7.8. In classical electromagnetism the Lorentz force on a particle moving at $\vec{v}$ in a magnetic field $\vec{B}$ is proportional to $\vec{v} \times \vec{B}$. The magnetic field is related to a vector potential $\vec{A}$ by $\vec{B} = \nabla \times \vec{A}$, so the force is proportional to $\vec{v} \times (\nabla \times \vec{A})$. Show

that the velocity dependent force you obtained in Exercise 7.8 has exactly this form, with h_{0j} playing the role of the vector potential. For this reason the force is often called a gravitomagnetic force. Having no classical analog it is peculiar to relativity and has been measured in satellite experiments (Everitt 2015; Adler 2000).

Chapter 8

Curved Space and Gravity



Abstract In this chapter we return to mathematics and study curvature in a Riemann space. Einstein's general relativistic field equations of gravity follow in an intuitive way from a study of the Riemann tensor and the geometric view of gravity.

8.1 Curved Space and the Riemann Tensor

When we studied vectors and tensors in Part II we often referred to curved 2-surfaces, depending on geometrical intuition for the meaning of curvature. Now however we must deal with the more sophisticated idea of a general curved space, because in general relativity gravity is described by a curved 4-dimensional spacetime; this is the natural outcome of the discussion of the last section on the geometric view of gravity (Misner 1973; Adler 1975; Schutz 2009). Indeed, we have already had an example of how to handle the analysis and definition of curvature when we parallel displaced a vector around a triangle on a plane and on a spherical surface in Chap. 5.

Consider first the familiar 2 and 3-dimensional spaces of Euclidean geometry. We call such a space a *Euclidean space*; a Euclidean space is defined by the property that there is a coordinate system in which the metric is equal to the identity matrix everywhere; thus a Euclidean space has signature $(1, 1, \dots, 1)$. For Euclidean 3-space, for example, the metric in the special system is

$$g_{ij} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \text{Euclidean 3-space.} \quad (8.1)$$

We may of course describe the space with other coordinate systems, such as spherical.

Next recall the space of special relativity, Minkowski space, which is usually coordinatized with ct and Cartesian coordinates. This is similar to a Euclidean space, but is distinguished by the minus signs in the metric. A *pseudo-Euclidean* space is defined by the property that there is a coordinate system in which the metric is equal everywhere to a diagonal matrix with $+1$ or -1 on the diagonal. For example, Minkowski space has in the special system the Lorentz metric

$$g_{\mu\nu} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}, \quad \text{pseudo-Euclidean spacetime.} \quad (8.2)$$

As usual we may use other coordinates if desired.

In these definitions we used a form of the metric with $+1$ or -1 on the diagonal, which is convenient. However, it is clear that if there is a coordinate system in which the metric is merely *constant*, then the space must be Euclidean or pseudo-Euclidean, for by a linear transformation the constant metric could be put into one of the above forms according to the Cayley-Sylvester theorem, which we stated in Chap. 4 and discussed in Appendix 1 in Chap. 4 (Perlis 1952; Arfken 1970).

From our discussion of gravity viewed as a geometric phenomenon in Chap. 7 it is clear that a pseudo-Euclidean space cannot describe a gravitational field, for then we could find a coordinate system in which the metric was everywhere the Lorentz metric, so that the classical potential in (7.28), would vanish, hence no gravity. It is thus necessary that space-time differ from pseudo-Euclidean Minkowski space in a fundamental way in order to describe gravity.

Let us then ask the following interesting question: in an arbitrary coordinate system how can we determine if the space is Euclidean or perhaps pseudo-Euclidean? It is clearly not practical to try every coordinate transformation to see if a constant metric results. We wish to find a covariant and more useful method. We do this as follows: in the special coordinate system where the metric is constant the connections are zero everywhere, as is obvious from their definition, so that ordinary and covariant derivatives are equal. Thus, in that special coordinate system the following string of equalities for the derivatives of any vector field ξ^α holds true

$$\xi^\alpha_{;\beta;\gamma} = \xi^\alpha_{,\beta,\gamma} = \xi^\alpha_{,\gamma,\beta} = \xi^\alpha_{;\gamma;\beta}, \quad \text{so } \xi^\alpha_{;\beta;\gamma} - \xi^\alpha_{;\gamma;\beta} = 0. \quad (8.3)$$

The last equation is a tensor equation, which we have obtained by using a special coordinate system; it is thus valid in all coordinate systems. We have thus proved the following theorem:

Theorem 1 *If a space is Euclidean or pseudo-Euclidean then for any vector field the antisymmetric combination of second covariant derivatives $\xi^\alpha_{;\beta;\gamma} - \xi^\alpha_{;\gamma;\beta}$ vanishes.*

This is a very useful and powerful criterion for determining whether a space is Euclidean or pseudo-Euclidean. With a little algebraic manipulation it can be put into even more elegant and useful form. We state this as a theorem.

Theorem 2 *The combination of second derivatives $\xi^\alpha_{;\beta;\gamma} - \xi^\alpha_{;\gamma;\beta}$ can be expressed as a linear combination of the vector components ξ^α , specifically.*

$$\xi^\alpha_{;\beta;\gamma} - \xi^\alpha_{;\gamma;\beta} = R^\alpha_{\eta\beta\gamma} \xi^\eta. \quad (8.4)$$

The tensor $R^\alpha_{\eta\beta\gamma}$ is called the Riemann curvature tensor or simply the Riemann tensor; it is constructed from the connections and their derivatives and will be calculated and defined explicitly below.

We will prove Theorem 2 and define the Riemann tensor by direct algebraic manipulation. First denote the covariant derivative as

$$T^\alpha_{\beta} = \xi^\alpha_{;\beta} = \xi^\alpha_{,\beta} + \Gamma^\alpha_{\beta\eta} \xi^\eta. \quad (8.5)$$

Then by the definition of covariant tensor derivatives

$$\begin{aligned} \xi^\alpha_{;\beta;\gamma} &= T^\alpha_{\beta;\gamma} = T^\alpha_{\beta,\gamma} + \Gamma^\alpha_{\tau\gamma} T^\tau_{\beta} - \Gamma^\lambda_{\beta\gamma} T^\alpha_{\lambda} \\ &= [\xi^\alpha_{,\beta,\gamma} + \Gamma^\alpha_{\beta\eta,\gamma} \xi^\eta + \Gamma^\alpha_{\beta\eta} \xi^\eta_{,\gamma}] \\ &\quad + \Gamma^\alpha_{\tau\gamma} [\xi^\tau_{,\beta} + \Gamma^\tau_{\beta\eta} \xi^\eta] - \Gamma^\lambda_{\beta\gamma} T^\alpha_{\lambda}. \end{aligned} \quad (8.6)$$

Clearly $\xi^\alpha_{;\gamma;\beta}$ is given by the same expression with β and γ reversed. From (8.6) it is easy to write the difference; we find

$$\begin{aligned} \xi^\alpha_{;\beta;\gamma} - \xi^\alpha_{;\gamma;\beta} &= \Gamma^\alpha_{\beta\eta,\gamma} \xi^\eta - \Gamma^\alpha_{\gamma\eta,\beta} \xi^\eta \\ &\quad + [\Gamma^\alpha_{\beta\eta} \xi^\eta_{,\gamma} - \Gamma^\alpha_{\gamma\eta} \xi^\eta_{,\beta} + \Gamma^\alpha_{\tau\gamma} \xi^\tau_{,\beta} - \Gamma^\alpha_{\tau\beta} \xi^\tau_{,\gamma}] \\ &\quad + \Gamma^\alpha_{\tau\gamma} \Gamma^\tau_{\beta\eta} \xi^\eta - \Gamma^\alpha_{\tau\beta} \Gamma^\tau_{\gamma\eta} \xi^\eta. \end{aligned} \quad (8.7)$$

But the terms in the square bracket cancel, and we are left with

$$\begin{aligned} \xi^\alpha_{;\beta;\gamma} - \xi^\alpha_{;\gamma;\beta} &= R^\alpha_{\eta\beta\gamma} \xi^\eta, \\ R^\alpha_{\eta\beta\gamma} &\equiv \Gamma^\alpha_{\beta\eta,\gamma} - \Gamma^\alpha_{\gamma\eta,\beta} + \Gamma^\alpha_{\tau\gamma} \Gamma^\tau_{\beta\eta} - \Gamma^\alpha_{\tau\beta} \Gamma^\tau_{\gamma\eta}. \end{aligned} \quad (8.8)$$

This proves the theorem and defines the very important Riemann tensor. Notice that it is built with only the connections and their derivatives, and of course it does not depend on the vector ξ^η : it is a purely geometrical object in that it is constructed from only the metric tensor. Also note that it is indeed a tensor by (8.8) and the quotient theorem. The Riemann tensor may look a bit formidable at first since it is fourth rank and is composed of many terms, but its importance makes it worth study.

From the above Theorem 2 we may restate Theorem 1 in a beautiful new way.

Theorem 3 *A Euclidean or pseudo-Euclidean space has a zero Riemann tensor.*

This is now obvious, being merely a restatement of Theorem 1.

We finally have come to the definition of a curved versus a flat space. We call a space *flat* if the Riemann tensor is zero, and *curved* if the Riemann tensor is not zero. This clearly fits our needs for a general and useful definition of flat space, since we see by Theorem 3 that Euclidean 2 and 3-space are flat, as is the Minkowski space of special relativity.

The converse of Theorem 3 is also true, that if the Riemann tensor is zero we can find a coordinate system in which the metric tensor is globally constant. The proof is a bit tedious, so we will not give it here but refer the reader to Adler (1975). Instead we will devote our time to another interesting geometric property that is equivalent to flatness. We state this as Theorem 4.

Theorem 4 *We can set up a constant vector field (that is one with zero covariant derivative) by parallel displacement from some initial vector at an initial point if and only if the space is flat, that is if the Riemann tensor is zero.*

This is a very restrictive and perhaps surprising theorem. The proof involves doing the construction explicitly and is straight-forward. Begin with a vector V^α at some arbitrary point in the space, and parallel displace it along some curve C to a point labeled x^λ , to produce the field $V^\alpha(x^\lambda)$. If this is to be a unique and well-defined field then it cannot depend on which curve between the initial point and x^λ one uses; a different curve C' would do as well. The covariant derivative of this field must be zero by construction; this is easy to see, since by definition

$$V^\alpha{}_{;\beta} dx^\beta = [V^\alpha{}_{,\beta} + \Gamma^\alpha_{\beta\gamma} V^\gamma] dx^\beta = dV^\alpha + \Gamma^\alpha_{\beta\gamma} V^\gamma dx^\beta. \quad (8.9)$$

Since we set the field up by parallel displacement

$$dV^\alpha = -\Gamma^\alpha_{\beta\gamma} V^\gamma dx^\beta, \quad \text{so } V^\alpha{}_{;\beta} = 0. \quad (8.10)$$

Because of this we may express the ordinary derivative of the field in terms of connections as

$$V^\alpha{}_{,\gamma} = -\Gamma^\alpha_{\beta\gamma} V^\beta. \quad (8.11)$$

But for a well-defined field the order of the ordinary second derivatives does not matter, $V^\alpha{}_{,\gamma,\delta} = V^\alpha{}_{,\delta,\gamma}$, so from (8.11)

$$\begin{aligned} (\Gamma^\alpha_{\beta\gamma} V^\beta)_{,\delta} &= (\Gamma^\alpha_{\beta\delta} V^\beta)_{,\gamma}, \\ \Gamma^\alpha_{\beta\gamma,\delta} V^\beta + \Gamma^\alpha_{\beta\gamma} V^\beta{}_{,\delta} &= \Gamma^\alpha_{\beta\delta,\gamma} V^\beta + \Gamma^\alpha_{\beta\delta} V^\beta{}_{,\gamma}. \end{aligned} \quad (8.12)$$

Using (8.11) to simplify this we write

$$\Gamma^\alpha_{\beta\gamma,\delta} V^\beta - \Gamma^\alpha_{\beta\gamma} \Gamma^\beta_{\delta\tau} V^\tau = \Gamma^\alpha_{\beta\delta,\gamma} V^\beta + \Gamma^\alpha_{\beta\delta} \Gamma^\beta_{\gamma\tau} V^\tau, \quad (8.13)$$

and relabel indices to see that

$$[\Gamma^\alpha_{\gamma\beta,\delta} - \Gamma^\alpha_{\delta\beta,\gamma} + \Gamma^\alpha_{\delta\tau} \Gamma^\tau_{\gamma\beta} - \Gamma^\alpha_{\gamma\tau} \Gamma^\tau_{\delta\beta}] V^\beta = R^\alpha{}_{\beta\gamma\delta} V^\beta = 0. \quad (8.14)$$

Since the vector may have any value we see that the Riemann tensor must vanish; conversely, if the Riemann tensor vanishes the construction goes through, so the

theorem is proved. Since the field as we have constructed it is independent of the path used to parallel displace the vector to the desired point we say that the space is *integrable*.

We may summarize the results of this section by saying that the following three properties of a space are equivalent:

1. The space is *Euclidean* or *pseudo-Euclidean*, so there is a coordinate system in which the metric is constant. It may, if desired, be put into the Cayley-Sylvester canonical form with positive and negative ones on the diagonal.
2. The space is *flat*, or the Riemann tensor is zero.
3. The space is *integrable*, so we may set up a constant vector field by parallel displacement, with a covariant derivative equal to zero.

The integrability property is noteworthy: in a curved space we cannot set up a constant vector field. See Fig. 5.5 for an illustration on the surface of a sphere.

The vanishing of the Riemann tensor is a very useful characteristic indeed, and in fact will lead us to the field equations of general relativity. In the next section we will study the symmetries of this fourth rank tensor.

8.2 Symmetries of the Riemann Tensor

The Riemann tensor is the largest tensor we have encountered so far. It is extremely important in Riemann geometry and in general relativity. In 4-dimensions it has $4^2 = 256$ components. However there are a number of symmetries which reduce this to only 20 *independent* components. These symmetries are easy to derive if we make use of the special geodesic coordinate system in which the connections vanish. In order to study the symmetries we must first lower an index on the Riemann tensor as it is defined in (8.8), for tensors can have symmetry only among indices of the same type. Thus we will study in the geodesic system the totally covariant $R_{\alpha\beta\gamma\delta}$ (Kenyon 1990).

Note that although the connections are zero at some selected point note their derivatives are not zero. In the geodesic system the Riemann tensor as defined in (8.8) has only the first two terms, instead of all 4 terms. That is

$$R^\lambda{}_{\beta\gamma\delta} = \Gamma^\lambda_{\beta\gamma,\delta} - \Gamma^\lambda_{\beta\delta,\gamma}, \quad \text{geodesic system.} \quad (8.15)$$

There is yet another simplification in the geodesic system. Since the connections vanish the ordinary derivatives of the metric tensor are equal to the covariant derivatives. But the covariant derivatives of the metric are zero by the Ricci theorem. Thus all the first derivatives of the metric tensor vanish at the selected point in the geodesic system. This is true for both the covariant and contravariant versions of the metric tensor. (Note that the second derivatives do not in general vanish at the selected point.) Because of this we can write out (8.15) as

$$\begin{aligned}
R^\lambda{}_{\beta\gamma\delta} &= \frac{1}{2}g^{\lambda\tau}(g_{\gamma\tau,\beta} + g_{\beta\tau,\gamma} - g_{\beta\gamma,\tau})_{,\delta} - \frac{1}{2}g^{\lambda\tau}(g_{\delta\tau,\beta} + g_{\beta\tau,\delta} - g_{\beta\delta,\tau})_{,\gamma} \\
&= \frac{1}{2}g^{\lambda\tau}(g_{\gamma\tau,\beta,\delta} - g_{\beta\gamma,\tau,\delta} - g_{\delta\tau,\beta,\gamma} + g_{\beta\delta,\tau,\gamma}), \quad \text{geodesic system.} \quad (8.16)
\end{aligned}$$

Thus we may lower an index to obtain the fully covariant Riemann tensor as

$$R_{\alpha\beta\gamma\delta} = \frac{1}{2}(g_{\gamma\alpha,\beta,\delta} - g_{\beta\gamma,\alpha,\delta} - g_{\delta\alpha,\beta,\gamma} + g_{\beta\delta,\alpha,\gamma}), \quad \text{geodesic system.} \quad (8.17)$$

This is now in a form where the symmetries are transparent.

The following symmetries follow by simply writing out the four terms of the Riemann tensor from (8.17):

$$R_{\alpha\beta\gamma\delta} = -R_{\alpha\beta\delta\gamma} \quad \text{antisymmetry in last pair of indices,} \quad (8.18a)$$

$$R_{\alpha\beta\gamma\delta} = -R_{\beta\alpha\gamma\delta} \quad \text{antisymmetry in first pair of indices,} \quad (8.18b)$$

$$R_{\alpha\beta\gamma\delta} = R_{\gamma\delta\alpha\beta} \quad \text{symmetry in interchange of index pairs.} \quad (8.18c)$$

There is one more symmetry for the 4-dimensional case; this is easily verified also by writing out all the terms using (8.17),

$$R_{0123} + R_{0231} + R_{0312} = 0. \quad (8.19)$$

This completes the algebraic symmetries. We emphasize that the symmetries have been obtained in the special geodesic coordinate system at the selected point, but a symmetry property of a tensor holds in any coordinate system, so the symmetries are generally true.

There is also a set of symmetries on the derivatives of the Riemann tensor that is easy to derive using the geodesic coordinate system. From the definition of the Riemann tensor in (8.8) we may differentiate it with respect to the coordinates. The right side of the defining equation will have two connection second derivative terms and four terms which involve the connections and their first derivatives. In the geodesic system the last four terms are clearly zero and we have thus

$$R^\alpha{}_{\eta\beta\gamma;\mu} = \Gamma^\alpha_{\beta\eta,\gamma,\mu} - \Gamma^\alpha_{\eta\gamma,\beta,\mu}. \quad (8.20)$$

But since this is the geodesic system the ordinary derivatives are the same as the covariant derivatives, so

$$R^\alpha{}_{\eta\beta\gamma;\mu} = \Gamma^\alpha_{\beta\eta,\gamma,\mu} - \Gamma^\alpha_{\eta\gamma,\beta,\mu}. \quad (8.21)$$

It follows from this that the following permuted combination is zero

$$R^\alpha_{\eta\beta\gamma;\mu} + R^\alpha_{\eta\gamma\mu;\beta} + R^\alpha_{\eta\mu\beta;\gamma} = 0, \quad (8.22)$$

which we see by writing out all the terms using the connections and their symmetry. These are called the Bianchi identities. As with the algebraic symmetries we have obtained the Bianchi identities in the geodesic coordinate system in which the connections vanish at the selected point, but they are tensor symmetries and thus hold in all coordinate systems. They will prove useful in obtaining the Einstein gravitational field equations.

8.3 The Einstein Equations for the Gravitational Field in Vacuum

There is a convincing heuristic path that leads from classical gravity to the field equations of general relativity (Adler 1975). Recall that classical gravity may be viewed in geometric terms if we relate the metric to the classical potential by (7.23), which we repeat here

$$g_{00} = 1 + \frac{2\phi}{c^2}, \quad \text{geometry} \leftrightarrow \text{classical gravity}. \quad (8.23)$$

From the discussion of the Riemann tensor we see moreover that the absence of a gravitational field corresponds to a zero Riemann tensor, for then there is a coordinate system in which the metric is Lorentz and the gravitational potential in (8.23) must vanish; that is $\phi = 0$ everywhere, and all the second derivatives vanish, $\phi_{,i,j} = 0$. Thus

$$R_{\alpha\beta\gamma\delta} = 0 \leftrightarrow \phi_{,i,j} = 0, \quad \text{absence of gravity}. \quad (8.24)$$

Indeed, using the correspondence in (8.23) we can make the above correspondence more explicit. As before we take the classical potential divided by c^2 to be very small and time independent, so the metric is nearly Lorentz. Working to lowest order we may express components of the Riemann tensor in terms of the classical potential; from the definition (8.8) we have

$$R^i_{0j0} = \Gamma^i_{0j,0} - \Gamma^i_{00,j} = -\Gamma^i_{00,j} = -\frac{1}{2}h_{00,i,j}, \quad (8.25)$$

where we have made use of the time independence of the metric and the connections, and have used (8.23) to calculate the connection. Now using (8.23) and (8.25) we obtain the important approximate relation

$$R^i_{0j0} = -\frac{1}{c^2}\phi_{,i,j} \quad \text{classical limit}. \quad (8.26)$$

The path to the field equations is now clear. The condition (8.24) that the Riemann tensor be zero corresponds to flat space and no gravity. If we weaken the condition on the classical potential from $\phi_{,i,j} = 0$ by summing over $i = j$ we get $\phi_{,i,i} = 0$, Laplace's equation, which is the correct equation for the classical potential in vacuum! Thus we are led to contract the Riemann tensor in exactly the same way and postulate for the gravitational field in vacuum,

$$\begin{aligned} R^\alpha{}_{\mu\alpha\nu} &= R_{\mu\nu} = 0, \quad \text{vacuum field equations,} \\ R_{\mu\nu} &\equiv \Gamma_{\beta\nu,\mu}^\beta - \Gamma_{\mu\nu,\beta}^\beta + \Gamma_{\tau\mu}^\beta \Gamma_{\beta\nu}^\tau - \Gamma_{\tau\beta}^\beta \Gamma_{\mu\nu}^\tau. \end{aligned} \quad (8.27)$$

The contracted Riemann tensor defined in (8.27) is called the Ricci tensor.

The Ricci tensor has several interesting properties. First it might seem that there are 6 different ways to contract the Riemann tensor, but the symmetries discussed in the last section imply that the different ways either give zero or the same result up to a sign. That is, the Ricci tensor is really the only independent contraction of the Riemann tensor. Secondly the Ricci tensor is symmetric, as may easily be shown (see Exercise 8.5). Thus it has only 10 independent components in 4 dimensions, so the field equation (8.27) are a set of 10 partial differential equations, the right number to determine the 10 components of the symmetric metric tensor.

There is another equivalent form for the field equation (8.27) that is mathematically interesting and will prove useful when we study the gravitational field in nonvacuum regions of space, that is where there is matter and energy present. This involves a tensor with zero divergence known as the Einstein tensor.

To get the alternative form we first calculate the divergence of the Ricci tensor, that is $R^\alpha{}_{\eta;\alpha}$. Using the Bianchi identities (8.22) we raise an index to find

$$R^{\alpha\eta}{}_{\beta\gamma;\delta} + R^{\alpha\eta}{}_{\gamma\delta;\beta} + R^{\alpha\eta}{}_{\delta\beta;\gamma} = 0. \quad (8.28)$$

Then we contract α with β and η with γ to get

$$R^{\alpha\eta}{}_{\alpha\eta;\delta} + R^{\alpha\eta}{}_{\eta\delta;\alpha} + R^{\alpha\eta}{}_{\delta\alpha;\eta} = 0, \quad \text{or} \quad R^\eta{}_{\eta;\delta} - R^\alpha{}_{\delta;\alpha} - R^\eta{}_{\delta;\eta} = 0. \quad (8.29)$$

Next we denote the contracted Ricci tensor, or Riemann scalar, as $R = R^\eta{}_\eta$, and relabel indices in (8.29) to obtain the divergence of the Ricci tensor

$$R^\alpha{}_{\delta;\alpha} = \frac{1}{2} R^\eta{}_{\eta;\delta} = \frac{1}{2} R_{;\delta}, \quad \text{or} \quad R^{\mu\nu}{}_{;\nu} = \frac{1}{2} g^{\mu\nu} R_{;\nu}. \quad (8.30)$$

Having obtained the divergence of the Ricci tensor in (8.30) we may define a tensor with a zero divergence, called the Einstein tensor, as

$$G^{\mu\nu} = R^{\mu\nu} - \frac{1}{2} g^{\mu\nu} R. \quad (8.31)$$

The zero divergence follows trivially from (8.30). Note also that the Einstein tensor is clearly symmetric.

A simple theorem is the key to the new form of the field equations.

Theorem 5 *The Einstein tensor is zero if and only if the Ricci tensor is zero. The proof is the simple Exercise 8.6.*

Thus the field equation (8.28) may also be written as

$$G^{\mu\nu} = 0 \quad \text{zero divergence form of field equations.} \quad (8.32)$$

The fact that the Einstein tensor has zero divergence will prove very useful when we add matter and energy to the picture, especially in the study of cosmology.

In this section we have tried to motivate the field equations (8.27) or (8.32) heuristically as the natural covariant generalization of classical gravity. Of course the test of their correctness is to solve them for physically interesting cases and compare the result to experiment, as we will do in the next chapters.

8.4 The Non-vacuum Field Equations

We have so far considered only gravity in free space, that is in vacuum. Now we want to obtain the field equations in the presence of matter or energy, such as in the interior of a star or in the large-scale universe. In classical theory this involves going from the Laplace equation to the Poisson equation (7.6). That is

$$\nabla^2 \phi = 0 \quad \xrightarrow{\text{vacuum}} \quad \nabla^2 \phi = 4\pi G\rho, \quad \rho = \text{mass density.} \quad (8.33)$$

That is, in classical theory we simply place a quantity representing matter on the right side of the equation. We will do precisely the same thing for general relativity. We will take the field equations for vacuum (8.32) and replace the zero on the right side with an object that represents the mass and energy density in space.

$$G^{\mu\nu} = 0 \quad \xrightarrow{\text{vacuum}} \quad G^{\mu\nu} = CT^{\mu\nu}, \quad T^{\mu\nu} = \text{energy-momentum.} \quad (8.34)$$

The tensor on the right side is the source of the gravitational field. It is called the *energy-momentum* tensor for reasons that will become apparent when we consider some special cases; C is a constant to be determined, but we expect it to be proportional to Newton's constant G .

The field equations (8.34) are so general as to not mean much yet, since we have not discussed the nature of the energy-momentum tensor. There are two properties that the energy-momentum tensor must have, however. First it must be symmetric since the Einstein tensor is symmetric. Second it must have zero divergence, since the Einstein tensor has zero divergence, as we discussed in Sect. 8.3. That is

$$T^{\mu\nu}{}_{;\nu} = 0. \quad (8.35)$$

Of course the theory has been set up so that this must be true, and later in this section and in Part IV we will study the energy-momentum tensor of a fluid to see what the divergence condition means physically.

Before we further consider the physical meaning of the energy-momentum tensor let us do a bit of tensor algebra and write the field equations (8.34) in yet another equivalent way that will prove useful in finding the classical limit. We write the Einstein tensor according to its definition (8.31) in terms of the Ricci tensor and substitute it into the field equations (8.34) to get the field equations explicitly in terms of the Ricci tensor,

$$G^{\mu\nu} = R^{\mu\nu} - \frac{1}{2}g^{\mu\nu}R = CT^{\mu\nu}. \quad (8.36)$$

Next we contract this to find a relation between the Riemann scalar and the contracted energy-momentum tensor $T = T^{\nu}{}_{\nu}$

$$R^{\nu}{}_{\nu} - \frac{1}{2}g^{\nu}{}_{\nu}R = R - 2R = -R = CT. \quad (8.37)$$

From this we may write the field equation (8.36) in terms of the Ricci tensor rather than the Einstein tensor,

$$R_{\mu\nu} = C\left(T_{\mu\nu} - \frac{1}{2}g_{\mu\nu}T\right). \quad (8.38)$$

We may use either the Einstein tensor or the Ricci tensor in writing the field equations, depending on convenience. The above form (8.38) will be useful in the next section.

In practice there are a number of ways to obtain the energy-momentum tensor of a given type of material. In this section we will consider only the simplest, that for an idealized material that is often called “dust.” In Part IV we will discuss a more general fluid describing the contents of the universe on a large scale.

Dust is defined as a fluid having only a mass-energy density and a flow velocity field u^{α} but no pressure or other properties, as shown in Fig. 8.1. There is one obvious symmetric second rank tensor we can build from the density and flow velocity, which is

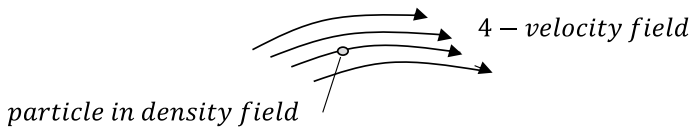


Fig. 8.1 At any point in spacetime the dust fluid has only a density and a velocity

$$T^{\alpha\beta} = \rho u^\alpha u^\beta, \quad u^\beta = \frac{dx^\beta}{ds} \quad \text{velocity along flow lines in dust fluid.} \quad (8.39)$$

Note that we here use a dimensionless 4-velocity u^β equal to the usual 4-velocity over c . We will refer to the dust tensor often. For now we will study it mainly in its classical limit, that is for zero or weak gravity and low velocities. This will let us verify that the field equations lead to the classical Poisson equation (8.33), and also let us evaluate the proportionality constant C . The task is easy since we have already done most of the needed calculations in Chap. 7 when we studied the link between geometry and gravity.

As in Chap. 7, where we studied the classical limit, we assume that the metric is the Lorentz metric plus a small time-independent perturbation,

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}, \quad h_{\mu\nu} \ll 1. \quad (8.40)$$

Moreover for consistency we must assume that the density of the material producing the field is small, that is of order $h_{\mu\nu}$, and also assume that it moves slowly. Then the flow velocity field is approximately that of special relativity with negligible 3-velocity,

$$u^\beta = \frac{dx^\beta}{ds} \cong (1, 0, 0, 0). \quad (8.41)$$

The energy-momentum tensor (8.39) then has only the 0,0 component, which is equal to the mass density, and the right side of the field equations in the form (8.38) is

$$C \left(T_{\mu\nu} - \frac{1}{2} g_{\mu\nu} T \right) = \frac{1}{2} C \rho \delta_{\mu\nu}. \quad (8.42)$$

The Ricci tensor on the left side of the field equation (8.38) is easy to obtain as we did in Chap. 7. The connections are of order $h_{\mu\nu} \ll 1$ so the second two terms of the Ricci tensor, defined in (8.27), may be neglected so we have approximately

$$R_{\mu\nu} = \Gamma_{\beta\nu,\mu}^\beta - \Gamma_{\mu\nu,\beta}^\beta. \quad (8.43)$$

Consider first the $\mu = \nu = 0$ component. Since the metric is time independent the connections are also time independent, and the Ricci tensor 0,0 component is easily obtained from (7.21)

$$R_{00} = -\Gamma_{00,j}^j = -\frac{1}{2} h_{00,j,j} = -\frac{1}{2} \nabla^2 h_{00}, \quad j = 1, 2, 3. \quad (8.44)$$

We now have explicit approximate expressions for both sides of the field equation (8.38) for $\mu = \nu = 0$. We substitute and obtain

$$\nabla^2 h_{00} = -\rho C. \quad (8.45)$$

But we know from Chap. 7 that the metric perturbation must be related to the classical potential by (7.23)

$$\phi = \frac{1}{2}c^2 h_{00}. \quad (8.46)$$

Thus we obtain

$$\nabla^2 \phi = -\frac{c^2}{2} C \rho. \quad (8.47)$$

We do indeed get the Poisson equation in the classical limit, and by comparison with Poisson's equation for classical gravity (8.33) we find the value of the constant C to be

$$C = -8\pi G/c^2, \quad \text{energy-momentum tensor using mass density.} \quad (8.48)$$

This completes our task: we have shown that the field equations of relativity reduce to the classical Poisson equation for the Newtonian gravitational potential, and the constant in the field equations is determined in (8.48). Note that the constant C is negative; this is the price we pay for our choice of the overall metric sign as we discussed in Part I. See also Exercise 8.10.

It is often more convenient to use an energy-momentum tensor with units of energy density, in which case the constant contains an additional factor of $1/c^2$,

$$C = -8\pi G/c^4, \quad \text{energy-momentum tensor using energy density.} \quad (8.49)$$

An important comment is in order concerning the energy-momentum tensor, which is the source of gravity. We noted that the Einstein field equations force it to be symmetric and have a zero divergence. The zero divergence condition means generally that the source is conserved. This is most easily shown for the simplest case of the dust tensor (8.39). For slowly moving material the 4-velocity is $u^\alpha = (1, \vec{v}/c)$ and the zero-divergence condition for $\mu = 0$ is

$$T^{0\nu}{}_{;\nu} = T^{00}{}_{;0} + T^{0k}{}_{;k} = \frac{1}{c} \left(\frac{\partial \rho}{\partial t} + \nabla \cdot \rho \vec{v} \right) = 0. \quad (8.50)$$

The last expression implies that mass is conserved: the time change of mass density is balanced by the mass flowing out of a small volume. It is an elegant facet of general relativity that the Einstein equations imply conservation of the source, whatever it might be. We will return to this in later chapters on cosmology when we discuss the energy-momentum tensor of a perfect fluid.

8.5 The Intrinsic Signature of Gravity

Recall our discussion of the equivalence principle. We concluded that because a uniform gravitational field is equivalent to acceleration of the reference frame it may be transformed away. Thus the intrinsic signature of gravity is its *non-uniformity* and not the presence of a force (Misner 1973) (see Figs. 7.2–7.4). This can now be seen in a very clear light. If we have a sufficiently large lab and sufficiently accurate equipment we may be able to detect the difference between gravitational forces in different parts of the lab. The force difference may be written as

$$dF^i = F^i_{,k} dx^k = -\phi_{,i,k} dx^k, \quad \text{signature of gravity: tidal forces.} \quad (8.51)$$

These are called tidal forces because they do indeed give rise to the tides on earth. That is, the intrinsic signature of Newtonian gravity is the non-vanishing of the second derivatives of the potential. But by the correspondence between the Riemann tensor and the potential derivatives in (8.27) we see that the corresponding signature of the gravitational field in relativity is

$$R^\alpha_{\beta\gamma\delta} \neq 0, \quad \text{signature of gravity: curved spacetime.} \quad (8.52)$$

The intrinsic signature of gravity is that the Riemann tensor is nonzero or space is curved. This is an invariant signature since if the Riemann tensor is nonzero in one coordinate system it is nonzero in all coordinate systems.

Let us summarize the general relativistic viewpoint on the equivalence principle and the intrinsic signature of gravity:

1. To the extent that the gravitational field is *uniform* over some small region of spacetime it is equivalent to an accelerated system. The gravitational force may thus be *transformed away*. This is clearly a *local* and thus approximate correspondence.
2. To the extent that the gravitational field *varies* over the relevant region of spacetime it corresponds to curvature of spacetime. The Riemann tensor field may not be transformed away. This is clearly a nonlocal and intrinsic characterization of the gravitational field that distinguishes it from the effects of acceleration.

As we have seen the equivalence principle as stated by Einstein leads to the idea of a geometric theory of gravity, and to some deep insights and correct predictions such as the deflection of light by gravity and the redshift of light in a gravitational field. It has played an important role in the development of general relativity and continues to elucidate problems concerning electromagnetic effects and quantum effects in a gravitational field. There is much more that could be said about the equivalence principle; there are at least 3 versions of it, of which we have only used the first, usually called the weak equivalence principle (WEP) or the universality of free fall. The reader is invited to pursue more deeply in the references the other versions and interpretations, as well as related experimental tests (Will 1993, 2014; Zee 1989).

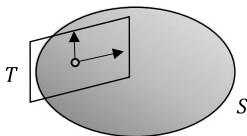


Fig. 8.2 The 2-surface S and the tangent plane T at point P , showing the special orthogonal coordinate systems

After these brief comments on the equivalence principle and its various forms and interpretations it is well to note a cautionary remark by Nordtvedt, that “Principles are for when you do not yet have a theory.”

Appendix 1: Tangent Spaces

Consider a 2-surface S imbedded in 3-dimensional Euclidean space. Intuition tells us that if S is reasonably smooth at some point P then there will be a flat plane T which coincides with it. The two spaces will be quite similar in a small region near P , as shown in Fig. 8.2. We call T the tangent plane at P . We emphasize that the two spaces S and T are different spaces which closely coincide (or osculate) only at the point P .

We can view this relation in the context of Appendix 2 in Chap. 4 and Appendix 1 in Chap. 5; there we showed that there exists a coordinate system in S for which the metric has the Cayley-Sylvester canonical form and vanishing first derivatives, so the connections are zero. The axes in this coordinate system are orthogonal and it clearly coincides closely with a global Cartesian coordinate system in the tangent plane T , as shown in Fig. 8.2. Notice the additional interesting fact that in the special coordinate system in S the covariant derivatives are the same as ordinary derivatives since the connections vanish at P .

This relation between a curved 2-dimensional Riemann space and a flat tangent plane can be generalized to higher dimensions and any signature. In general relativity theory a curved Riemann space, analogous to S , corresponds to a gravitational field. The space analogous to the tangent plane T is a flat Lorentz space and the coordinates may be taken to be the Minkowski coordinates; special relativity holds in this tangent Lorentz space, which coincides locally with the curved Riemann space. There is a gravitational field in the curved space, while there is none in the tangent Lorentz space.

Appendix 2: The Riemann Tensor as a 6 by 6 Matrix

There is an elegant way to view the Riemann tensor as a matrix in which the number of *independent* components becomes quite clear. It is also useful in classifying space-times (Petrov 1969). Think of the first pair of indices $\alpha\beta$ as a single index A . Since

the Riemann tensor is antisymmetric in this pair only 6 values of the pair occur (Adler 1975)

$$\begin{array}{ll} \text{tensor indices } \alpha\beta = & 23 \ 31 \ 12 \ 01 \ 02 \ 03 \\ \text{matrix indices } A = & 1 \ 2 \ 3 \ 4 \ 5 \ 6 \end{array} \quad (8.53)$$

Similarly for the pair γ, δ we may associate a 6 valued matrix index B . This allows us to think of the Riemann tensor as a 6 by 6 matrix R_{AB} . But the symmetry (8.18c) means that the matrix is symmetric in A and B so it has at most 21 independent components. The final symmetry in (8.19) is one more relation on the components and reduces the number of independent components to 20, much less than the total of 256.

Exercises

- 8.1 How many independent components does the Riemann tensor have in two dimensions, three dimensions, and four dimensions?
- 8.2 Consider a metric in two dimensions with coordinates x, y that has the special form $ds^2 = dx^2 + G^2(x)dy^2$. Show that one of the components of the Riemann tensor is

$$R^1{}_{212} = G \left(\frac{d^2 G}{dx^2} \right).$$

Obtain all the nonzero components from this one. The metric which we will use for cosmology will be analogous to this. See also Exercise 8.7.

- 8.3 What is the Riemann tensor for the 2-dimensional surface of a sphere? What is it for the surface of a cylinder? (Is Exercise 8.2 any help?)
- 8.4 What is the Riemann scalar for the surface of a sphere? Is this a surprise?
- 8.5 Prove that the Ricci tensor is symmetric.
- 8.6 Prove Theorem 5, that the Einstein tensor is zero if and only if the Ricci tensor is zero.
- 8.7 Consider a 4-dimensional spacetime with a particularly simple metric form,

$$ds^2 = (dx^0)^2 - g_{ik} dx^i dx^k, \quad i = 1, 2, 3,$$

where the g_{ik} are independent of the time marker x^0 . That is the 4-space contains a 3-space in a simple way. How are the 4-space connections related to the 3-connections?

- 8.8 For the metric of Exercise 8.7, what is the relation between the Riemann tensor in 4-space and that in 3-space? What is the relation between the Ricci tensor in 4-space and that in 3-space? What is the relation between the Riemann scalar in 4-space and that in 3-space?
- 8.9 Show that the gravity-free pseudo-Euclidean space of special relativity is a solution of the Einstein equations in vacuum. (This is as easy as it sounds.)

- 8.10 The sign of the constant C in the field equation (8.36) is negative according to (8.48). This is a drawback of our choice of the overall metric sign in Part I. Think through the development of the field equations in Chap. 8 and convince yourself that all is in order with either sign of the metric.

Chapter 9

Spherically Symmetric Gravitational Fields



Abstract This chapter begins with a derivation of the Schwarzschild solution, the single most important result of general relativity theory. It describes the gravitational field of a spherically symmetric body such as the sun. As an application of the Schwarzschild solution the orbits of planets and the deflection of star light can be obtained, and the comparison of these with observation gives strong evidence that the theory is correct.

9.1 The Schwarzschild Solution

We now turn to a study of the full nonlinear theory and obtain its best-known exact solution, that of Schwarzschild (1916). Because of its importance in physics and in history we will do this in considerable detail. The Einstein field equations for free space in (8.27) are a set of 10 partial differential equations. We repeat them explicitly here

$$R_{\mu\nu} = \Gamma_{\beta\nu,\mu}^{\beta} - \Gamma_{\mu\nu,\beta}^{\beta} + \Gamma_{\tau\mu}^{\beta} \Gamma_{\beta\nu}^{\tau} - \Gamma_{\tau\beta}^{\beta} \Gamma_{\mu\nu}^{\tau} = 0. \quad (9.1)$$

The first two terms contain second derivatives of the metric tensor, and there are many terms containing the metric and its first derivatives scattered about. We therefore have a set of equations that look a bit formidable. We cannot merely stare at them and write a solution, but instead must ponder the physical context and set the problem up cleverly to find solutions. The solution of Schwarzschild for the field of a spherically symmetric body is a beautiful example of this. It was obtained only about a year after Einstein first presented his vacuum field equations in 1915 (Einstein 1915, 1923; Schwarzschild 1916). It is certainly the most important solution in general relativity since it represents the exterior field of the sun and other stars (Misner 1973; Adler 1975).

It is natural to use spherical coordinates for the problem. In the absence of gravity the appropriate metric is

$$ds^2 = c^2 dt^2 - (dr^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\varphi^2). \quad (9.2)$$

This is simply the metric of flat space time, that is the Minkowski space of special relativity, expressed in spherical space coordinates. To describe the gravitational field it must be modified. We already know approximately what the modification must be, for the relation between the metric and the Newtonian field in (7.23) tells us that for a body of mass M

$$g_{00} = 1 + \frac{2\phi}{c^2} = 1 - \frac{2GM}{c^2 r}. \quad (9.3)$$

Thus it is clear that we should allow g_{00} to be a function of the radial coordinate. Moreover we may guess that since the field is spherically symmetric we must allow g_{11} to also be a function of r . We thus look for a solution of the form

$$ds^2 = e^{\nu(r)} c^2 dt^2 - e^{\lambda(r)} dr^2 - r^2 (d\theta^2 + \sin^2 \theta d\varphi^2). \quad (9.4)$$

The use of exponential functions is entirely for future mathematical convenience. Thus we have guessed a simple form of metric with only two unknown functions of r , which must be chosen to satisfy the field equations. Note that we have allowed the angular part of the line element to remain in exactly the same form as the flat space line element. The simplicity of (9.4) is the key to the exact solution.

The next step in solving the field equations is to calculate the connections needed to write out the Ricci tensor. This is straight-forward if we use the shortcut discussed in Chap. 5, and is left as an exercise for the reader. Most of the connections are zero, and the task is correspondingly easy, giving

$$\begin{aligned} \Gamma_{10}^0 &= \Gamma_{01}^0 = \frac{\nu'}{2}, & \Gamma_{00}^1 &= \frac{\nu'}{2} e^{\nu-\lambda}, & \Gamma_{11}^1 &= \frac{\lambda'}{2}, \\ \Gamma_{22}^1 &= -e^{-\lambda} r, & \Gamma_{33}^1 &= -e^{-\lambda} r \sin^2 \theta, & \Gamma_{12}^2 &= \Gamma_{21}^2 = \frac{1}{r}, \\ \Gamma_{33}^2 &= -\sin \theta \cos \theta, & \Gamma_{13}^3 &= \Gamma_{31}^3 = \frac{1}{r}, & \Gamma_{23}^3 &= \Gamma_{32}^3 = \cos \theta. \end{aligned} \quad (9.5)$$

Here the prime denotes a derivative with respect to r . Next we obtain the metric determinant, which is needed to calculate the contracted connection using (6.28), and find

$$\log \sqrt{|g|} = \frac{\nu + \lambda}{2} + 2 \log r + \log |\sin \theta|. \quad (9.6)$$

(Recall that $|g|$ is taken as the absolute value of the determinant as we noted in Chap. 4.)

We are now ready to write out the field equations in terms of the coordinates and the unknown functions ν and λ .

First we consider the Einstein equation for $\mu = \nu = 0$. From the Ricci tensor in (9.1) and the contracted connection in (6.18) we have

$$R_{00} = \left(\log \sqrt{|g|} \right)_{,0,0} - \Gamma_{00,\alpha}^\alpha + \Gamma_{\tau 0}^\beta \Gamma_{\beta 0}^\tau - \Gamma_{00}^\tau \left(\log \sqrt{|g|} \right)_{,\tau} = 0. \quad (9.7)$$

Many of the terms in this component are zero. From (9.5) and (9.6) we find it reduces to

$$\begin{aligned} R_{00} &= -\Gamma_{00,1}^1 + 2\Gamma_{00}^1 \Gamma_{01}^0 - \Gamma_{00}^1 \left(\log \sqrt{|g|} \right)_{,1} \\ &= -\left(\frac{1}{2} v' e^{v-\lambda} \right)' + \left(\frac{1}{2} v'^2 e^{v-\lambda} \right) - \left(\frac{1}{2} v' e^{v-\lambda} \right) \left(\frac{v' + \lambda'}{2} + \frac{2}{r} \right) \\ &= -\frac{e^{v-\lambda}}{2} \left(v'' + \frac{1}{2} v'^2 - \frac{1}{2} \lambda' v' + \frac{2}{r} v' \right) = 0 \end{aligned} \quad (9.8)$$

Thus the $\mu = \nu = 0$ equation distills down to

$$v'' + \frac{1}{2} v'^2 - \frac{1}{2} \lambda' v' + \frac{2}{r} v' = 0. \quad (9.9)$$

In the same manner we work out the $\mu = \nu = 1$ term of the field equations and find

$$\begin{aligned} R_{11} &= \left(\log \sqrt{|g|} \right)_{,1,1} - \Gamma_{11,\alpha}^\alpha + \Gamma_{\tau 1}^\beta \Gamma_{\beta 1}^\tau - \Gamma_{11}^\tau \left(\log \sqrt{|g|} \right)_{,\tau} \\ &= \left(\log \sqrt{|g|} \right)_{,1,1} - \Gamma_{11,1}^1 + \Gamma_{01}^0 \Gamma_{01}^0 + \Gamma_{11}^1 \Gamma_{11}^1 + \Gamma_{21}^2 \Gamma_{21}^2 + \Gamma_{31}^3 \Gamma_{31}^3 \\ &\quad - \Gamma_{11}^1 \left(\log \sqrt{|g|} \right)_{,1} = \frac{1}{2} \left(v'' + \frac{1}{2} v'^2 - \frac{1}{2} \lambda' v' - \frac{2}{r} \lambda' \right) = 0, \\ \text{so } v'' + \frac{1}{2} v'^2 - \frac{1}{2} \lambda' v' - \frac{2}{r} \lambda' &= 0 \end{aligned} \quad (9.10)$$

Equations (9.9) and (9.10) suffice to find the unknown functions. They are ordinary second order differential equations.

Subtracting (9.10) from (9.9) we see that

$$(v + \lambda)' = 0, \quad v + \lambda = \text{const.} \quad (9.11)$$

Differential equations generally require *boundary conditions*. In this problem the appropriate boundary condition is quite obvious: we ask that the metric be that of gravity-free flat space at a large distance from the origin; that is the line element should approach (9.2). This in turn means that the two unknown functions ν and λ must both approach zero. Thus the constant in (9.11) must be zero, and

$$\lambda = -\nu. \quad (9.12)$$

We now substitute this into (9.10) and find the following equation for ν

$$v'' + v'^2 + \frac{2}{r} v' = 0. \quad (9.13)$$

This is sufficiently simple that we may solve it in the traditional way, that is we inspect it, make a transformation or two, and guess a solution. We first transform to a new function f ,

$$v = \log f, \quad v' = \frac{f'}{f}, \quad v'' = \frac{ff'' - f'^2}{f^2}, \quad (9.14)$$

so that (9.13) simplifies to

$$f'' + \frac{2}{r}f' = 0. \quad (9.15)$$

Note that $f = g_{00}$ from (9.4). The solution to (9.15) is obviously a power, so we try $f = r^n$ and find

$$n^2 + n = 0, \quad \text{thus,} \quad n = 0, \quad n = -1. \quad (9.16)$$

Thus the solution is

$$g_{00} = e^v = f = A - \frac{2m}{r}. \quad (9.17)$$

Here A and $2m$ are constants of integration, and it only remains to determine them. This is easy because we know the metric at large distance from the body in (9.3). We see thereby from the classical limit

$$A = 1, \quad m = \frac{GM}{c^2}. \quad (9.18)$$

Let us now collect the results in (9.17), (9.12), and (9.4) and write the Schwarzschild line element as

$$ds^2 = \left(1 - \frac{2GM}{c^2 r}\right) c^2 dt^2 - \left(1 - \frac{2GM}{c^2 r}\right)^{-1} dr^2 - r^2(d\theta^2 + \sin^2\theta d\varphi^2). \quad (9.19)$$

The reader should never forget this result. It is certainly the most well-known and important solution of the theory.

A few words about the line element (9.19) are in order. Note that we have used only two of the 10 field equations. It is straight-forward to verify that the other 8 equations are satisfied by the Schwarzschild metric (9.19), and this is left as an exercise.

The parameter m is a constant of integration, which we related to the Newtonian mass M and gravitational constant G using the classical limit relation (9.3). It has the dimension of a distance, and is called the *geometric mass*. For the sun it is about 1.47 km. The quantity $2m$ is called the Schwarzschild radius, and is a key quantity in black hole physics, which we will discuss in the next chapter. It is very important to understand that the solution (9.19) is valid only in vacuum outside the spherically

symmetric body. For the sun it is valid for radii greater than the solar radius, which is about 10^6 km. For smaller radii we must solve a different problem. See Exercises 9.10 and 9.11 and Chap. 10.

Finally it is worth noting that the approximate classical limit for g_{00} in (9.3) is, in this case, exact. This is an accident, due to the choice of coordinates, and has no deep meaning. In Appendix 1 we will obtain the metric in other coordinates, for which (9.3) is only approximate.

Digression 9.1 The *Birkhoff Theorem* states that the Schwarzschild solution is the *unique* solution to the vacuum field equations for the exterior of a spherically symmetric body, given the Minkowski space boundary condition (Birkhoff 1923; Misner 1973). This means that the assumption of time independence of the metric that we made above is in fact not necessary.

The coordinates used in the metric (9.19) are naturally called Schwarzschild coordinates. Another coordinate system, called spatially isotropic coordinates, is often used in the linearized theory and in discussions of the observational tests of general relativity. See Appendix 1 for this form. We will return to it in Chap. 11 (Will 1993, 2014).

9.2 Orbit of a Planet

Schwarzschild's solution is the key to studying the motion of the planets in the solar system. Since this is such an important problem we will work out in detail the orbit of a planet around the sun in this section, and will see that it is very nearly an ellipse, as in classical theory, but with a small change peculiar to relativity. Our solution will follow very closely the classical Kepler problem (Goldstein 1980). The reader need not know the classical theory to follow our solution, but it will be easier and more transparent if he does. The various transformations and tricks that we will use in this section are almost the same as those used in the classical problem.

We know from Chap. 7 that the equations of motion for a particle in spacetime are the Euler-Lagrange equations for a Lagrangian function constructed from the line element s ,

$$L = \left(1 - \frac{2m}{r}\right) c^2 \dot{t}^2 - \left(1 - \frac{2m}{r}\right)^{-1} \dot{r}^2 - r^2 \dot{\theta}^2 - r^2 \sin^2 \theta \dot{\phi}^2. \quad (9.20)$$

The dot denotes differentiation with respect to the line element s . Recall also that the numerical value of this Lagrangian function is 1. It is easy to work out the Euler-Lagrange equations for the coordinates t , θ , ϕ , and we find

$$\frac{d}{ds} \left[\left(1 - \frac{2m}{r} \right) \dot{t} \right] = 0, \quad \text{thus} \quad \left(1 - \frac{2m}{r} \right) \dot{t} = \ell = \text{const.}, \quad (9.21a)$$

$$\frac{d}{ds} [r^2 \dot{\theta}] = r^2 \sin \theta \cos \theta \dot{\phi}^2, \quad (9.21b)$$

$$\frac{d}{ds} [r^2 \sin^2 \theta \dot{\phi}] = 0, \quad \text{thus} \quad r^2 \sin^2 \theta \dot{\phi} = h = \text{const.} \quad (9.21c)$$

Notice that because the metric is independent of time and azimuthal angle these equations are rather simple. We could also write down the Euler-Lagrange equation for r , but it is simpler to recall that the value of the Lagrangian is 1, and use that in place of the Euler-Lagrangian equation for r ; in fact it is the first integral of that Euler-Lagrange equation. Thus we have

$$1 = \left(1 - \frac{2m}{r} \right) c^2 \dot{t}^2 - \left(1 - \frac{2m}{r} \right)^{-1} \dot{r}^2 - r^2 \dot{\theta}^2 - r^2 \sin^2 \theta \dot{\phi}^2. \quad (9.21d)$$

The first step in solving the system (9.21) is physically motivated; we expect the orbit to lie in a plane because of the spherical symmetry of the problem. Thus we try to find a solution in which the body moves in the equatorial plane $\theta = \pi/2$. We substitute this into the angular equations (9.21b) and (9.21c) and find that (9.21b) is identically satisfied, and the other equations simplify to

$$\left(1 - \frac{2m}{r} \right) \dot{t} = \ell, \quad (9.22a)$$

$$r^2 \dot{\phi} = h, \quad (9.22b)$$

$$1 = \left(1 - \frac{2m}{r} \right)^{-1} c^2 \ell^2 - \left(1 - \frac{2m}{r} \right)^{-1} \dot{r}^2 - \frac{h^2}{r^2}. \quad (9.22c)$$

The next step in the solution is to ask not for the coordinates r and φ as functions of S , but instead for the orbit radius expressed as a function $r(\varphi)$. Then

$$r' = \frac{dr}{d\varphi} = \frac{\dot{r}}{\dot{\varphi}}, \quad \text{thus} \quad \dot{r} = r' \dot{\varphi} = \frac{r' h}{r^2}. \quad (9.23)$$

We place this in (9.22c) to get

$$1 - \frac{2m}{r} = c^2 \ell^2 - \frac{h^2 r'^2}{r^4} - \frac{h^2}{r^2} \left(1 - \frac{2m}{r} \right). \quad (9.24)$$

Our next manipulation is to use the inverse of the radius, $u = 1/r$, rather than the radius r . Then

$$r' = -\frac{u'}{u^2}, \quad (9.25)$$

and the radial equation (9.24) becomes

$$u'^2 + u^2 = \left(\frac{c^2 \ell^2 - 1}{h^2} \right) + \frac{2mu}{h^2} + 2mu^3. \quad (9.26)$$

Next we perform another trick and differentiate (9.26) to get a second order equation; we do this because the second order equation is a close analog of the classical equation and easy to solve for the orbit. Thus

$$2u'u'' + 2uu' = \frac{2mu'}{h^2} + 6mu^2u', \quad \text{thus} \quad u'' + u = \frac{m}{h^2} + 3mu^2. \quad (9.27)$$

We will see that the first term on the right gives the classical solution, an elliptic orbit, and the second term gives a small relativistic correction.

Let us pause to consider the special case of circular orbits, which is a fair approximation for planets in the solar system. Take the radius to be a constant $r = r_c$, so that (9.27) becomes

$$\frac{1}{r_c} = \frac{m}{h^2} + \left(\frac{3m}{r_c} \right) \frac{1}{r_c}, \quad \text{circular orbit.} \quad (9.28)$$

For the sun the geometric mass m is of order 1 km , which is very much smaller than any planetary orbit, so $3m/r_c$ is a small dimensionless quantity; moreover it is thus clear that the first term m/h^2 on the right side of (9.28) must be about $1/r_c$.

Having established the approximate relative size of terms let us return to the general equation (9.27) and rewrite it as

$$u'' + u = A + \frac{\varepsilon}{A} u^2, \quad A \equiv \frac{m}{h^2}, \quad \varepsilon \equiv 3mA \ll 1. \quad (9.29)$$

where A has the dimension of inverse distance, and ε is small and dimensionless.

Solution of (9.29) is a nice exercise in perturbation theory. We expand the solution as a power series in the small parameter ε and work to first order in ε . This gives immediately the zeroth and first order equations

$$u = u_0 + \varepsilon u_1, \quad u''_0 + u_0 = A, \quad u''_1 + u_1 = \frac{u_0^2}{A}. \quad (9.30)$$

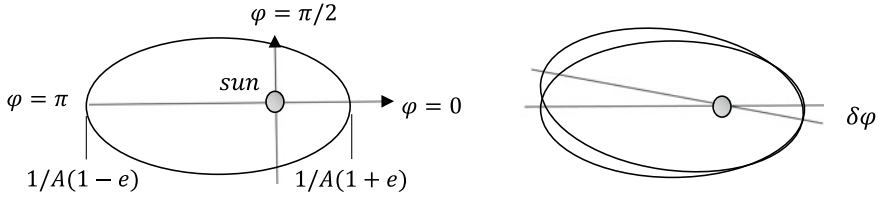


Fig. 9.1 The elliptic orbit of a planet in classical mechanics, and on the right the slowly precessing elliptic orbit in relativity

The zeroth order equation gives the classical orbit, as expected. The solution is

$$\begin{aligned} u_0 &= A + B \cos \varphi, \quad \text{so } r_0 = \frac{1}{u_0} = \frac{1}{A + B \cos \varphi} \\ &= \frac{1}{A(1 + e \cos \varphi)}, \quad e \equiv \frac{B}{A}. \end{aligned} \quad (9.31)$$

This is the famous elliptical solution of the classical problem of planetary orbits. Figure 9.1 shows the shape of the classical orbit.

The minimum radius, or *perihelion*, occurs at $\varphi = 0$, and the maximum radius, or *aphelion*, occurs at $\varphi = \pi$; these radii to zeroth order are

$$r_{0\min} = \frac{1}{A(1+e)}, \quad r_{0\max} = \frac{1}{A(1-e)}. \quad (9.32)$$

The positive parameter e is a measure of the non-circularity of the orbit and is called the *eccentricity*. Its value is less than 1 for any elliptic orbits, and is much less than 1 for the planets of the solar system.

The effect of relativity will be seen in the first order equation in (9.30). With the solution of the zeroth order equation in hand we may write the first order equation as

$$\begin{aligned} u_1'' + u_1 &= \frac{u_0^2}{A} = A + 2B \cos \varphi + \frac{B^2}{A} \cos^2 \varphi \\ &= \left(A + \frac{B^2}{2A} \right) + 2B \cos \varphi + \frac{B^2}{2A} \cos 2\varphi. \end{aligned} \quad (9.33)$$

This equation is relatively easy to solve. Since it is linear we may split the solution up into 3 terms, with each term being the solution of a simpler equation. That is we set $u_1 = u_{1a} + u_{1b} + u_{1c}$ and solve the three equations

$$u_{1a}'' + u_{1a} = A + \frac{B^2}{2A}, \quad u_{1b}'' + u_{1b} = 2B \cos \varphi, \quad u_{1c}'' + u_{1c} = \frac{B^2}{2A} \cos 2\varphi. \quad (9.34)$$

These three may be solved by inspection. The solutions are

$$u_{1a} = A + \frac{B^2}{2A}, \quad u_{1b} = B\varphi \sin \varphi, \quad u_{1c} = \frac{B^2}{6A} \cos 2\varphi. \quad (9.35)$$

Note that we have not included the homogeneous solutions of (9.34) in the above since they are already included in the zeroth order solution (9.31). Now we collect the results, the zeroth order solution in (9.31) and the first order solution in (9.35), to obtain

$$u = \left[A + \varepsilon \left(A + \frac{B^2}{2A} \right) \right] + B[\cos \varphi + \varepsilon \varphi \sin \varphi] - \left[\varepsilon \frac{B^2}{6A} \cos 2\varphi \right]. \quad (9.36)$$

Looking at these 3 terms we see that the first term corresponds to a slightly larger orbit than the classical orbit, the third corresponds to a small doubly periodic bulge in the orbit, and the second is the most interesting in that it may grow large for large angles; it is called a secular term. We thus ignore the third term, call the constant term $\tilde{A}$, and rewrite (9.36) as

$$u = \tilde{A} + B[\cos \varphi + \varepsilon \varphi \sin \varphi], \quad \tilde{A} = A + \varepsilon \left(A + \frac{B^2}{2A} \right). \quad (9.37)$$

To see the physical effect of the secular term we use the identity

$$\cos(1 - \varepsilon)\varphi = \cos \varphi \cos \varepsilon \varphi + \sin \varphi \sin \varepsilon \varphi = \cos \varphi + \varepsilon \varphi \sin \varphi, \quad (9.38)$$

and re-express the solution (9.37) as

$$u = \tilde{A} + B \cos(1 - \varepsilon)\varphi. \quad (9.39)$$

The physical behavior of the orbit is now clear. It is approximately an ellipse, but the period is not exactly 2π . It has now become

$$\frac{2\pi}{(1 - \varepsilon)} \cong 2\pi(1 + \varepsilon). \quad (9.40)$$

Thus successive perihelia and aphelia do not occur at the same place in the orbit, but *advance* by a small amount

$$\delta\varphi = 2\pi\varepsilon = 6\pi m\tilde{A}. \quad (9.41a)$$

The orbit is thus a slowly precessing ellipse as shown in Fig. 9.1.

A convenient approximate expression for the precession of a planet follows from (9.41a); for a planet in a nearly circular orbit we know from (9.32) that $\tilde{A}$ is approximately the same as the inverse of the orbital radius r_c , and from (9.36) $\tilde{A}$ is approximately the same as A , so (see Exercise 9.8)

$$\delta\varphi \cong 6\pi \frac{m}{r_c}. \quad (9.41b)$$

To evaluate the precession more precisely for an elliptic orbit we need to evaluate more accurately the constant $\tilde{A}$ in (9.41a). Astronomers routinely measure a planet's eccentricity e and its *semimajor axis*, which is defined according to Fig. 9.1 as

$$a = \frac{1}{2} \left[\frac{1}{A(1+e)} + \frac{1}{A(1-e)} \right] = \frac{1}{A(1-e^2)}, \quad \text{so } A = \frac{1}{a(1-e^2)}. \quad (9.42)$$

Thus we can express the precession in terms of the parameters a and e , which can be found in astronomy textbooks, as

$$\delta\varphi = \frac{6\pi}{1-e^2} \frac{m}{a}. \quad (9.43)$$

This can now be conveniently compared with the observations of planets.

The orbit precession is most easily measured for the planet Mercury since the semimajor axis is least for Mercury, and also since the perihelion position is accurately measurable for Mercury's fairly eccentric orbit. Indeed it was well-known before the development of general relativity that Mercury's orbit precesses by about $43''$ per century more than predicted by classical theory (Le Verrier 1859; Newcomb 1895). Equation (9.43) gives about $43''$ per century for Mercury, which provided a historically important verification of general relativity theory in its earliest days. Indeed, Einstein himself calculated the perihelion shift and was aware of this agreement (Einstein 1923). We will return to the question of the observational verification of general relativity in more detail in Sect. 9.4.

9.3 Deflection of Light

We have already noted how the equivalence principle predicts that light will fall in a gravitational field. In this section we will explicitly calculate the orbit of a light ray as it passes by a star such as the sun. The calculation is much like that for the orbit of a planet, so we can rely heavily on the analysis of the previous section.

We first consider the nature of the orbit of a light ray or photon. In special relativity the path of light is characterized by a null line element or $ds^2 = 0$. We naturally carry this over to general relativity as a fundamental assumption. We also carry over the geodesic motion of a particle and assume that light also follows a geodesic. Thus we make the well-justified assumption that light follows a *null geodesic*. Recall that the geodesic equation may use as an invariant curve parameter the line element ds or a parameter proportional to it, $dp = ds/\alpha$, where α is a constant. Recall also that the function L that plays the role of a Lagrangian for the motion of bodies has the value 1 if ds is used as a curve parameter,

$$L = g_{\mu\nu} \frac{dx^\mu}{ds} \frac{dx^\nu}{ds} = \frac{ds^2}{ds^2} = 1, \quad \text{particles.} \quad (9.44)$$

If instead $dp = ds/\alpha$, is used the Lagrangian has the value

$$L = g_{\mu\nu} \frac{dx^\mu}{dp} \frac{dx^\nu}{dp} = \frac{ds^2}{dp^2} = \alpha^2, \quad \text{particles or light.} \quad (9.45)$$

It is thus clear how we may analyze null geodesics: we take the limit $\alpha^2 \rightarrow 0$ in the above so as to force the line element to zero, the function L to zero, and use a parameter dp proportional to ds in the geodesic equation.

Let us do this explicitly for the Schwarzschild metric. Most of the analysis of the preceding section goes through unchanged for the null geodesic, except that the curve parameter is dp instead of ds , and the left side of (9.22c) is 0 and not 1. We then repeat the previous planetary orbit analysis and obtain an equation like (9.27) except that the constant term on the right side is absent, so that

$$u'' + u = 3mu^2. \quad (9.46)$$

We now solve this as in the planetary problem. Let the distance of closest approach of the photon to the body be $r_c = 1/u_c$, which we take to be much greater than m , and define as in the planetary problem a small dimensionless parameter $\varepsilon = 3m/r_c = 3mu_c$. Then the orbit equation reads

$$u'' + u = \varepsilon \frac{u^2}{u_c}. \quad (9.47)$$

As before we solve this by perturbation theory, and set

$$u = u_0 + \varepsilon u_1, \quad (9.48)$$

so that we have from (9.47) a zeroth order and a first order equation

$$u_0'' + u_0 = 0, \quad u_1'' + u_1 = \frac{u_0^2}{u_c}. \quad (9.49)$$

The zeroth order equation is trivial, and the desired solution with arbitrary constant C is

$$u_0 = C \sin \varphi, \quad \text{or} \quad r_0 \sin \varphi = \frac{1}{C} = r_c. \quad (9.50)$$

This describes an undeflected straight line path as shown in Fig. 9.2, just as we should expect.

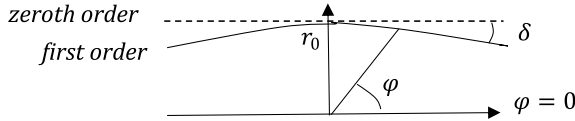


Fig. 9.2 The path of the photon or light ray showing the deflection

To obtain the deflection caused by the gravitational field we solve the first order equation in (9.49), using the zeroth order solution from (9.50),

$$u_1'' + u_1 = \frac{u_0^2}{u_c} = \frac{\sin^2 \varphi}{r_c} = \frac{1}{2r_c} - \frac{\cos 2\varphi}{2r_c}. \quad (9.51)$$

Splitting up the solution into two parts, $u_1 = u_{1a} + u_{1b}$, we turn this into the two equations

$$u_{1a}'' + u_{1a} = \frac{1}{2r_c}, \quad u_{1b}'' + u_{1b} = -\frac{\cos 2\varphi}{2r_c}. \quad (9.52)$$

The solutions to these two differential equations are easily checked to be

$$u_{1a} = \frac{1}{2r_c}, \quad u_{1b} = \frac{\cos 2\varphi}{6r_c}. \quad (9.53)$$

Now we collect results from the zeroth order (9.50) and the first order (9.53) to get

$$u = \frac{\sin \varphi}{r_c} + \frac{\varepsilon}{2r_c} \left[1 + \frac{\cos 2\varphi}{3} \right]. \quad (9.54)$$

To calculate the angle δ in Fig. 9.2 we observe that the radius is infinite and $u = 0$ for $\varphi = -\delta$, and moreover δ is taken to be very small. That is, from Fig. 9.2 and (9.54), we have to lowest order

$$\sin \varphi = -\delta, \quad \cos 2\varphi = 1. \quad (9.55)$$

Then (9.54) becomes for this case

$$0 = \frac{-\delta}{r_c} + \frac{\varepsilon}{2r_c} \left[\frac{4}{3} \right] \quad \text{so} \quad \delta = \frac{2}{3}\varepsilon = \frac{2}{3} \left(\frac{3m}{r_c} \right) = \frac{2m}{r_c}. \quad (9.56)$$

Finally, from Fig. 9.2, the total deflection is twice this, or

$$\Delta = \frac{4m}{r_c} = \frac{4GM}{c^2 r_c}, \quad \text{Einstein deflection.} \quad (9.57)$$

For starlight just grazing the edge of the sun this angle is about $1.75''$ of arc. This was measured for the first time during an eclipse in 1919, and the observed deflection was found to be in agreement with the relativity prediction to about 30% (Von Klüber 1960). It was a major triumph for general relativity because the observation came after the theoretical prediction. It signaled the end of the long era in which Newtonian gravitational theory was considered essentially perfect. We will discuss this further in the next section.

9.4 Observational Tests of General Relativity

Some brief comments on observational tests of general relativity are in order at this point. The literature on this subject is now vast so we will mention only a few useful sources: the book of Will is a standard reference, and has a wealth of detail, including an update chapter (Will 1993). See also the useful “living review” by Will available on the internet (Will 2014). Also on the internet the Wikipedia article is useful and generally up to date (Wiki TGR).

Much of the work on testing weak gravity now uses the isotropic coordinates and the parametrized post Newtonian (PPN) system discussed in Appendix 1.

There are *three classic tests* of general relativity proposed by Einstein. The first classic test is the gravitational redshift which we discussed in connection with the equivalence principle. Since the gravitational redshift can be derived from the equivalence principle it cannot serve as a test of the field equations and the full theory, but is nevertheless important since the equivalence is a conceptual cornerstone of general relativity.

Early attempts to measure the redshift using stars such as white dwarfs were not accurate enough to be satisfactory and the effect was not well-verified for stars until the 1950s (Hetherington 1980). However terrestrial tests by Pound et al. in the 1950s and 1960s using the Mossbauer effect definitively agreed with the theory to about 1% (Pound 2000). A later experiment, called Gravity Probe A, used a clock in a rocket boosted to about 10^4 km, and yielded a result in agreement with theory to about one part in 10^4 (Vessot 1980). Finally it is interesting to note that the GPS (Global Positioning System) must be corrected for red shift effects or it would be in error by many meters (Ashby 2003); thus the red shift is now continually being tested and verified by everyone using the GPS!

The perihelion shift of Mercury is the second classic test (Adler 1975; Will 2014). The anomalous precession of Mercury’s perihelion was well-known as early as 1859, long before general relativity (Leverrier 1859; Newcomb 1895). One early proposed solution to the problem was to postulate a new planet orbiting very close to the sun, called Vulcan, which was never detected. Because this anomaly was already known the calculation by Einstein was not a prediction, but was a very strong indication that general relativity was correct. There was once some dispute about the amount of precession contributed by the quadrupole moment of the sun, but this has largely been resolved, with the quadrupole contribution now believed to be only

about $0.03''$ per century. The presently accepted observational value of the precession as determined by radar measurements, about $42.98''$ per century, agrees with general relativity to about a part in 10^3 (Will 2014; Wiki ND). The precessions of other solar system planets have now been measured, as has the precession in both the Hulse-Taylor system and the double pulsar, which we will discuss below; all agree with the predictions of general relativity.

The deflection of starlight by the sun is the third classic test. As we noted previously the deflection was first measured for starlight during an eclipse of the sun in 1919, and an agreement of about 30% with theory was found (Von Klüber 1960). More recently radio sources have been used for the test, so that much more accurate and dependable results have been obtained. The agreement is now better than about a part in 10^3 (Kenyon 1990; Will 2014).

A further basic solar system test has been added to the three classic tests; light or radar signals passing near the sun are delayed by the gravitational field, an effect which is easily calculable and amounts to some hundreds of microseconds depending on the geometry of the experiment (Adler 1975). With radar reflected from planets and signals from planetary probes this effect has been accurately measured by Shapiro et al., and agrees with theory to better than 1% (Shapiro 1971; Kenyon 1990; Will 2014).

The equivalence principle has been subjected to many diverse tests since 1900, and the most accurate tests to date indicate that the inertial and gravitational masses of a body are equal to better than a part in 10^{12} (Will 2014). This is impressive, but there are various plans for space tests of the equivalence principle to an accuracy of about 10^{18} using satellites (Wiki STEP).

Some of the most important tests of general relativity outside the solar system involve binary pulsar systems. One is PSR1913 + 1916, which is a pulsar in a short period orbit, about 8 h, around an unseen companion, presumably a neutron star. It is widely called the Hulse-Taylor system after its discoverers. Because the system is small the relativistic effects are large. With only timing of the pulsar signals all of the orbital tests discussed above have been done for the system and are consistent with general relativity. Most important, the orbit has been observed to decay, which indicates an energy loss to gravitational radiation, and which agrees with relativity to within a few percent; we will discuss the process in Chap. 11. This is a most impressive result, and before the direct detection of waves by LIGO it was the only observational evidence for gravitational waves (Kenyon 1990; Will 2014). More recently a binary system of two pulsars, PSR J0737-3039A, has been discovered. Since it is smaller than the Hulse-Taylor system it promises to provide even more accurate tests (Burgay 2012).

All of the above tests involve weak gravity, in that the deviation of the metric components from those of flat space is small, less than about a part in 10^6 for the solar system. These tests are without question very important, but it is also important to test the theory for strong gravity, that is where the metric components deviate from flat space of order unity. We will discuss strong gravity in the next chapter.

Appendix 1: Isotropic Form of the Metric, Eddington Parameters

For some purposes it is useful to transform the Schwarzschild solution (9.19) to spatially isotropic form, that is with the space part of the metric equal to a multiple of the 3-dimensional flat space line element (Adler 1975). We can obtain such a spatially isotropic form by transforming from the Schwarzschild radius r to a new radial coordinate ρ defined by

$$r = \rho \left(1 + \frac{m}{2\rho} \right)^2. \quad (9.58)$$

The metric in the new coordinates is then obtained as

$$ds^2 = \left(1 - \frac{m}{2\rho} \right)^2 \left(1 + \frac{m}{2\rho} \right)^{-2} c^2 dt^2 - \left(1 + \frac{m}{2\rho} \right)^4 d\vec{x}^2. \quad (9.59)$$

This isotropic form will occur in the linearized theory that we will develop in Chap. 11.

Eddington suggested that the above isotropic metric be expanded for distances r large compared to m , where the field is weak, and written in terms of dimensionless parameters as (Eddington 1988; Adler 1999),

$$ds^2 = \left(1 - \alpha \frac{2m}{\rho} + \beta \frac{2m^2}{\rho^2} \dots \right) c^2 dt^2 - \left(1 + \gamma \frac{2m}{\rho} \dots \right) d\vec{x}^2. \quad (9.60)$$

The parameters have the values $\alpha = \beta = \gamma = 1$ for general relativity theory and are known as the Eddington parameters. The series metric (9.60) is clearly a rather general form for the metric far from a spherically symmetric body. An observational test of general relativity for weak gravity, such as in the solar system, can then be thought of as a measurement of how the three Eddington parameters compare with unity.

Since the constant m which appears in the metric represents the mass of the central body the parameter α may be absorbed into it, which is equivalent to taking $\alpha \equiv 1$. This is consistent so long as no independent non-gravitational determination of the central body mass is possible.

The Eddington parameters may be viewed as a book-keeping tool for tracking which terms in the metric are responsible for some gravitational effect, for example the deflection of starlight by the sun. Alternatively, they may be viewed as numbers which may not be equal to 1 if a metric theory other than general relativity is valid. In either case they provide a convenient way to express the results of experimental tests of gravity by giving values for the parameters. This parametrized approach has been extended to include many other parameters and has been highly developed under the name Parametrized Post-Newtonian theory or PPN (Will 1993).

In brief summary, solar system experiments and observations indicate that β can differ from unity by less than about a part in 10^3 and γ can differ from unity by less than about a part in 10^3 (Will 1993). General relativity is a well-tested theory.

Exercises

- 9.1 Verify the connections in (9.5) for the Schwarzschild metric in the form (9.4).
- 9.2 Verify the metric determinant function in (9.6).
- 9.3 Verify the field equations (9.9) and (9.10).
- 9.4 Check that the field equations $R_{22} = 0$ and $R_{33} = 0$ are satisfied by the Schwarzschild solution.
- 9.5 Check that the off-diagonal terms of the Ricci tensor are zero for the Schwarzschild form of metric (9.4). Thus the Einstein equations are entirely satisfied by the Schwarzschild solution.
- 9.6 What is the Schwarzschild radius for the sun? For the earth? For a proton? For a typical galaxy?
- 9.7 Show explicitly that the classical orbit given by (9.31) is an ellipse.
- 9.8 Calculate the numerical value for the precession of Mercury's orbit using (9.43) and data from an astronomy text. Also compare the eccentricity e of Mercury's orbit to that of the other planets.
- 9.9 Calculate the numerical value for the deflection of starlight just grazing the sun from (9.57).
- 9.10 Consider the region $r < 2m$ for the Schwarzschild metric. What is the sign of the 0,0 component and what is the sign of the 1,1 component? Is it thus clear that t cannot be interpreted as a time marker and r cannot be interpreted as a radial marker in this region? What happens if we reverse the interpretation of the two? There has been much written on appropriate coordinates in this region, the simplest and best known being called the Kruskal-Zekeres coordinates (Kruskal 1960; Adler 1975).
- 9.11 Do you foresee any problems in obtaining observational information from the interior Schwarzschild region $r < 2m$? See Chap. 11 for further discussion of this, and also (Adler 2005).
- 9.12 Use the references (Will 1993, 2014) and see how the three classic tests of relativity depend on the Eddington parameters, and also how the Shapiro time delay depends on the Eddington parameters.
- 9.13 Why is it that the Eddington expansion (9.60) includes quadratic terms in the time part of the metric but only includes linear terms in the space part of the metric?

Chapter 10

Black Holes and Gravitational Collapse



Abstract Black holes are one of the strangest predictions of relativity theory. In this chapter we study some properties of black holes and discuss how they are expected to be the end result of the collapse of some types of stars in the real universe. One extraordinary theoretical property of black holes is that they should radiate energy like a classical black body; this profound prediction connects classical general relativity with quantum theory, although the radiation has not yet been observed.

10.1 Schwarzschild Black Hole

A typical star like the sun is roughly spherically symmetric and has a geometric mass of about 10^6 km and a radius of about 10^6 km. Thus the Schwarzschild radius is deep within such a star as shown in Fig. 10.1.

Assuming it is approximately spherically symmetric we know the metric is the Schwarzschild metric for the exterior of such a star—but only the exterior. We have not yet studied the interior, which is an entirely different problem. For a typical star such as the sun the metric function $1 - 2m/r$ differs from 1 by less than about a part in 10^6 , so gravity is indeed weak. For a dense star it may become significantly less than 1 and gravity is strong. We will study the exterior of such a dense star in this section but will not discuss the interior until later.

Consider first the gravitational redshift of light from the surface of a small dense star, with its radius near to $2m$. The gravitational redshift of light from the stellar surface is given by (7.25), which we rewrite in terms of the frequency as

$$\frac{\nu_{ob}}{\nu_s} = \frac{\sqrt{g_{00}(s)}}{\sqrt{g_{00}(ob)}} = \frac{\sqrt{1 - 2m/r_s}}{\sqrt{1 - 2m/r_{ob}}}. \quad (10.1)$$

Here s refers to the stellar surface and ob refers to the observer, typically at a large distance from the surface. We see that for $r_s \rightarrow 2m$ the observed frequency goes to zero. Thus a photon emitted from a body at the surface loses all of its energy as it travels outwards. This means that light does not actually escape. Such a star

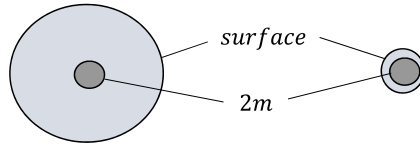


Fig. 10.1 A typical star with radius much greater than the Schwarzschild radius and a small dense star with a radius only slightly greater than the Schwarzschild radius

is invisible and is called a *black hole*, which is now a very well-known name. The black hole surface is referred to as an infinite redshift surface.

Let us study the behavior of light near a black hole in more detail; it is quite odd. For simplicity we consider light falling radially inward, so that $d\varphi = d\theta = 0$. Then from the fundamental postulate that the line element is null along the path of a photon the Schwarzschild metric (9.19) implies

$$ds^2 = \left(1 - \frac{2m}{r}\right)c^2 dt^2 - \left(1 - \frac{2m}{r}\right)^{-1} dr^2 = 0. \quad (10.2)$$

In this line element we may interpret the time t as that measured by an observer far outside the Schwarzschild radius. From (10.2) we may thus write the coordinate velocity of light as

$$v_c = \frac{dr}{dt} = \pm \left(1 - \frac{2m}{r}\right)c. \quad (10.3)$$

For light falling inward the minus sign applies. Of course this velocity is not the constant c since it is only the coordinate velocity, and coordinates are arbitrary markers of space and time location as we have stressed. Indeed the physical velocity which an observer measures is given by the physical distance interval in the r direction, which is $\sqrt{g_{11}}dr$, divided by the proper time interval, which is $\sqrt{g_{00}}dt$, so the physical velocity is indeed the absolute constant c ,

$$\frac{\sqrt{g_{11}}dr}{\sqrt{g_{00}}dt} = \pm \frac{1}{(1 - 2m/r)}(1 - 2m/r)c = \pm c. \quad (10.4)$$

We thus see that even though a local observer would measure the velocity of light to be c according to (10.4) light approaching $2m$ seems to slow and stop according to (10.3)! It is thus natural to ask if it would ever reach the Schwarzschild radius $2m$.

To answer this question we integrate (10.3) to get the coordinate time elapsed for the photon to go from a far point, labeled f , to a near point labeled n ,

$$c\Delta t = - \int_{r_f}^{r_n} \frac{dr}{\left(1 - \frac{2m}{r}\right)} = (r_f - r_n) + 2m \log\left(\frac{r_f - 2m}{r_n - 2m}\right). \quad (10.5)$$

As expected the photon takes an infinite time to reach the black hole surface where $r_n = 2m$. It is easy to show from (10.5) that when all the distances are near $2m$ the photon approaches the black hole asymptotically as

$$r_n - 2m = A \exp(-c\Delta t/2m), \quad A = \text{const.} \quad (10.6)$$

Let us repeat this analysis for a massive particle falling radially onto a black hole. We will find an interesting result; we will also apply the result later to the collapse of an idealized star with zero pressure, that is a dust star. The equations of motion were obtained when we studied the motion of a planet and are given in (9.22). For radial fall with $\dot{\theta} = \dot{\phi} = 0$ the relevant equations are

$$\left(1 - \frac{2m}{r}\right) \dot{t} = \ell, \quad (10.7a)$$

$$1 = \left(1 - \frac{2m}{r}\right)^{-1} c^2 \ell^2 - \left(1 - \frac{2m}{r}\right)^{-1} \dot{r}^2. \quad (10.7b)$$

We first evaluate the constant of integration ℓ . Suppose we drop the particle from rest at r_f so that (10.7b) gives, at that point,

$$1 = \left(1 - \frac{2m}{r_f}\right)^{-1} c^2 \ell^2, \quad \text{so} \quad 1 - c^2 \ell^2 = \frac{2m}{r_f}. \quad (10.8)$$

Then (10.7b) becomes

$$\dot{r}^2 = \frac{2m}{r} - \frac{2m}{r_f}. \quad (10.9)$$

Now we may solve for $r(t)$ using (10.7a) and (10.9). We get

$$\frac{dr}{dt} = \frac{\dot{r}}{\dot{t}} = \frac{\sqrt{2m/r - 2m/r_f}}{\sqrt{1 - 2m/r_f}} (1 - 2m/r)c, \quad (10.10)$$

and the time to fall from r_f to r_n is the integral

$$c\Delta t = -\sqrt{1 - 2m/r_f} \int_{r_f}^{r_n} \frac{dr}{\sqrt{2m/r - 2m/r_f}(1 - 2m/r)}. \quad (10.11)$$

The interesting part of the fall is near the black hole surface at $2m$ so we suppose for simplicity that the far radius r_f is much larger than $2m$, and the integral becomes simple,

$$c\Delta t = \frac{2}{3\sqrt{2m}} \left[r_f^{3/2} - r_n^{3/2} + 6m(\sqrt{r_f} - \sqrt{r_n}) \right] + 2m \log \left[\frac{(\sqrt{r_f} - \sqrt{2m})(\sqrt{r_n} + \sqrt{2m})}{(\sqrt{r_f} + \sqrt{2m})(\sqrt{r_n} - \sqrt{2m})} \right]. \quad (10.12)$$

We see that as $r_n \rightarrow 2m$ the time required becomes infinite, just as for the photon. The particle never reaches the black hole surface. It is easy to show from (10.12) that the particle approaches the black hole asymptotically exactly like a photon, or

$$r_n - 2m = A \exp(-c\Delta t/2m), \quad A = \text{const.} \quad (10.13)$$

Both the photon and the particle fall onto the black hole surface exponentially, and never quite reach it. (However see Exercise 10.7 and Sect. 19.6 on very small distances in physics!)

The above analysis gives the motion of the particle in terms of the coordinate time, $r(t)$. This is appropriate from the point of view of an observer far outside the black hole whose proper time is approximately equal to the Schwarzschild coordinate time. We may also analyze the motion in terms of the proper time of an observer falling with the particle towards the black hole, whose proper time is the arc length divided by c . For this we need only integrate (10.9) for $r(s)$,

$$\frac{dr}{ds} = -\sqrt{\frac{2m}{r}}, \quad \Delta s = \frac{2}{3\sqrt{2m}} (r_f^{3/2} - r_n^{3/2}). \quad (10.14)$$

As before we have taken the far point r_f to be much greater than $2m$. This time is totally different from the previous result (10.12). From the viewpoint of the observer falling with the particle it falls onto the black hole surface in a finite time

$$\Delta s_{\text{BH}} = \frac{2}{3\sqrt{2m}} (r_f^{3/2} - (2m)^{3/2}). \quad (10.15)$$

The behavior of the particle falling onto the black hole is shown in Fig. 10.2, both from the point of view of the distant observer using Schwarzschild time and also from the point of view of the observer falling with it onto the surface using his own proper time. The difference is infinitely large, which stems from the behavior of the

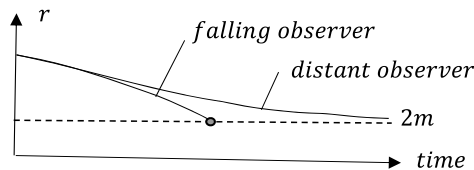


Fig. 10.2 Fall of a particle towards the surface of a Schwarzschild black hole as seen by a distant exterior observer and by an observer falling with the particle

metric function $1 - 2m/r$, which goes to zero at the surface. This is a profound difference characteristic of black holes!

As seen from the outside, where physicists live, light and particles falling onto a black hole would take an infinite amount of time to reach the surface; however, if one were to fall onto the surface of the black hole carrying a clock he would reach the surface in the finite time (10.15). For the falling observer the entire history of the external world would thus be seen to pass during his fall. This remarkable behavior has never been directly tested by a falling physicist, but observations of material falling onto a black hole from an accreting disk of matter are consistent with it.

As noted earlier we have not discussed the interior region, $r < 2m$. This is because the Schwarzschild coordinates simply do not work there. Indeed the signature changes from $(1, -1, -1, -1)$ to $(-1, 1, -1, -1)$ at the Schwarzschild radius, so t cannot be thought of as a time coordinate and r cannot be thought of as a radial coordinate (see Exercise 9.10). Moreover the point $r = 0$ should *not* be thought of as the “center” of the black hole, a fact which is not always appreciated. To study the interior one must use other coordinates that should be consistent with Schwarzschild coordinates outside but remain well-behaved inside the Schwarzschild radius. The best-known coordinates of this type are called the Kruskal Szekeres coordinates and serve their purpose quite well. See Exercise 10.9 and Kruskal (1960).

Another thing we should note about the Schwarzschild line element is that at the Schwarzschild radius the time term goes to zero while the radial term becomes infinite, but the product of the two in the determinant remains finite. Thus the 4-space volume element is well behaved at the Schwarzschild radius but the 3-space volume element is not.

We also emphasize that we have not yet discussed how a Schwarzschild black hole could form in the real universe that we observe, for example from a collapsing star. We will discuss this in more detail when we study the collapse of model stars in Sect. 10.4.

10.2 Null Surfaces

We have seen in the previous section that the black hole surface has some interesting properties. In particular the behavior of both light and particles is quite peculiar as they approach the surface from outside. In terms of the time used by external observers, such as physicists, neither light nor particles can reach the surface. Indeed the surface is special in a way that is independent of the choice of coordinates; the surface is called a null surface and we will see that it acts like a one-way membrane or horizon. In this section we will study the relation of a general surface in spacetime to the local light cone and obtain some elegant geometric results characterizing a null surface that are relevant to black holes.

Consider a smooth surface S in spacetime defined by

$$S: f(x^\alpha) = C = \text{const.} \quad (10.16)$$

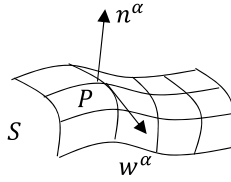


Fig. 10.3 The surface S with normal n^α and tangent vector w^α at P

The vector $n_\alpha = f_{,\alpha}$, the gradient of f , is normal to the surface since its inner product with any dx^α on the surface is zero,

$$n_\alpha dx^\alpha = f_{,\alpha} dx^\alpha = df = 0, \quad (10.17)$$

since f is constant on S . This is shown in Fig. 10.3.

At any point P on S we may find a coordinate system in which the metric is the Lorentz metric of special relativity according to the signature theorem of Chap. 4, and in that system the line element is

$$ds^2 = (dx^0)^2 - (dx^1)^2 - (dx^2)^2 - (dx^3)^2. \quad (10.18)$$

The local light cone is defined by $ds^2 = 0$ at P , or in the local Lorentz system

$$(dx^0)^2 - (dx^1)^2 - (dx^2)^2 - (dx^3)^2 = 0. \quad (10.19)$$

By a rotation in 3-space we can always place the x axis along the 3-vector part of the normal vector, so it takes the form

$$\begin{aligned} n^\alpha &= (n^0, n^1, 0, 0), & n_\alpha &= (n_0, -n_1, 0, 0), \\ n^2 &= n^\alpha n_\alpha = (n^0)^2 - (n^1)^2. \end{aligned} \quad (10.20)$$

Consider a tangent vector w^α to S at P , that is a vector lying along some dx^α . Since the normal and the tangent are orthogonal we see that

$$n_\alpha w^\alpha = n_0 w^0 - n_1 w^1 = 0, \quad \text{so} \quad \frac{w^0}{w^1} = \frac{n^1}{n^0}. \quad (10.21)$$

Thus the tangent vector may be written as

$$w^\alpha = \lambda (n^1, n^0, a, b), \quad (10.22)$$

where a , and b and λ are arbitrary real numbers. The norm of w is thus

$$\begin{aligned} w^2 &= w^\alpha w_\alpha = \lambda^2 [(n^1)^2 - (n^0)^2 - (a^2 + b^2)] \\ &= -\lambda^2 [n^2 + a^2 + b^2]. \end{aligned} \quad (10.23)$$

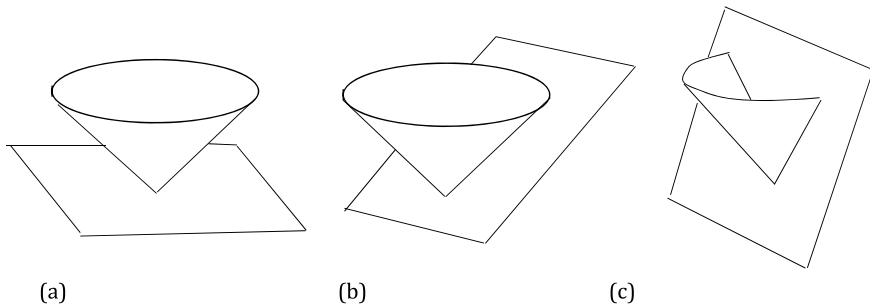


Fig. 10.4 The 3 cases of surface orientation with respect to the local light cone

This relation between the norms of the normal and tangent vectors leads to a beautiful geometric result with profound physical consequences.

Case I: n^α is *timelike*, so $n^2 > 0$. Then w^2 is negative from (10.23), that is w^α is spacelike and lies outside the light cone. There is thus no tangent vector which lies along the local light cone. The geometric situation is shown in Fig. 10.4a, with one space dimension ignored.

Case II: n^α is *null*, so $n^2 = 0$. Then w^2 is negative unless $a = b = 0$, in which case it is zero. There is thus one tangent vector direction which can lie along the local light cone. The geometric situation is shown in Fig. 10.4b.

Case III: n^α is *spacelike*, or $n^2 < 0$. Then w^2 may be positive or negative or zero. Thus there is a family of tangent vectors which lie on the local light cone. There is also a family of tangent vectors which lie inside the light cone. The geometric situation is illustrated in Fig. 10.4c.

The physical interpretation of the geometry in Fig. 10.4 is quite clear. Since massive particles have trajectories within the local light cone and photons have trajectories on the local light cone we see that physical objects can pass through a spacelike surface (Case III) in either direction, and can pass through a timelike surface (Case I) in only one direction. The null surface is the dividing or critical case; it is the configuration where one-way behavior begins, and we identify it as a *one-way membrane*.

A simple example of a null surface or one-way membrane may be taken from special relativity. The surface $ct = 0$ is timelike (has a timelike normal) and objects may pass only in the forward time direction. The surface $x = 0$ is space-like (has a spacelike normal) and objects may pass in either direction. The surface $ct = x$ is null (has a null normal) and objects may pass in only one direction; one tangent vector of the surface that lies on the local light cone is $w^\alpha = (1, 1, 0, 0)$.

A more interesting case is the black hole surface at $r = 2m$. A spherical surface in Schwarzschild coordinates has a normal

$$n_\alpha = (0, 1, 0, 0), \quad \text{so} \quad n^2 = g^{\alpha\beta} n_\alpha n_\beta = -\left(1 - \frac{2m}{r}\right). \quad (10.24)$$

Thus the spherical surface is spacelike outside the Schwarzschild radius, and becomes null on it. By the above geometric arguments we see that the surface of a Schwarzschild black hole is a null surface, and we therefore expect that *nothing from the interior could pass through it and emerge outside*, neither a particle nor light. The name black hole is thus appropriate. On the other hand objects may fall onto it in terms of their proper time or approach it asymptotically in terms of the external Schwarzschild time, as we discussed in the preceding section.

The above comments are based on classical relativity. If quantum effects are considered however the situation changes in an interesting way: a black hole may emit radiation as if it were a black body. We will study this later in Sect. 10.7.

For the Schwarzschild case that we have discussed the surface at $r = 2m$ is both an infinite redshift surface and a null surface. In the more general case of a black hole which is rotating these two surfaces are not the same, and we will discuss this further in Sect. 10.5 on the Kerr metric.

10.3 Stellar Evolution, Very Briefly

A typical star is born when a gas cloud, mainly of hydrogen, collapses under the influence of gravity. As it collapses it heats up as gravitational potential energy is converted into heat energy of the gas. When the temperature has risen sufficiently high a number of thermonuclear processes begin, such as the fusion of protons via the overall process $4p \rightarrow \text{He} + 2e^+ + 2\nu + \text{radiation}$. These release energy as heat and radiation, and the pressure due to the increasing temperature and radiation pressure stabilize the star against further collapse. It then becomes a stable energy emitting star for a relatively long time. We can discuss briefly and superficially the behavior of some stars at the end of their stable lifetime. See also the material on the death of stars in the free online textbook of Fraknoi (2016).

A *low mass* star like the sun emits radiation for billions of years until its hydrogen is depleted and the radiation pressure can no longer stabilize it. It then collapses into a denser and denser state, and may emit material from its surface as it does, called a planetary nebula. It finally becomes small and dense, about the size of the earth with a density roughly 10^6 times water. This is called a white dwarf. It is prevented from further collapse by the pressure of the degenerate electron gas in its dense interior, much as the electrons in a metal make the metal highly incompressible. Such white dwarfs are stable only if their mass is less than about 1.4 solar masses, which is called the *Chandrasekhar limit*.

A *medium mass* star will also emit radiation, but for a shorter time, until its hydrogen is depleted and radiation pressure can no longer stabilize it. Unlike a low mass star it may then eject large amounts of material and huge amounts of energy in a spectacular and complicated supernova explosion. The remnant left behind in such an explosion may have a mass greater than the Chandrasekhar limit of 1.4 solar masses; if that is the case it cannot be a white dwarf. In such a remnant the electrons may be absorbed by protons via inverse beta decay to form neutrons, and a neutron

star is formed. A neutron star is even smaller and denser than a white dwarf, of order 10 km in radius, and with roughly nuclear density, 10^{14} times water. Many neutron stars have been observed as pulsars and all have masses greater than or about the Chandrasekhar limit. Pulsars are neutron stars that may spin at rates of up to about 10^3 Hz and emit electromagnetic radiation in regularly spaced pulses. The quantity $2m/r$ may be of order 1/10 for a neutron star, indicating a much stronger gravitational field than occurs in the solar system. The Hulse-Taylor binary pulsar system PSR 1913 + 16 mentioned in Chap. 9 is a pulsar in orbit with a companion neutron star; the companion does not emit radiation in our direction (Hulse 1975).

A *very heavy* star will emit radiation for a still shorter time, and may also undergo a supernova explosion. If the core remnant of the explosion is sufficiently massive however it cannot form a neutron star. There is an upper limit to the mass of a neutron star, called the Tolman-Oppenheimer-Volkoff or TOV limit, analogous to the Chandrasekhar limit for a white dwarf (Tolman 1939). The numerical value of the TOV limit is not as precisely known as the Chandrasekhar limit, but it is thought to be roughly two or three solar masses. The uncertainty is due mostly to lack of knowledge of the equation of state of the neutron fluid and the effect of rotation in the star. For a stellar remnant heavier than the TOV limit the internal pressure cannot support it against gravity and it shrinks to smaller and smaller size, until it finally approaches the Schwarzschild radius. The collapse of the remnant towards the Schwarzschild radius is thus somewhat like the fall of a particle onto the surface of a black hole that we studied in a previous section; it continues forever as viewed by an outside observer and the surface approaches the Schwarzschild radius asymptotically. In the final stage of the collapse the light from the surface of the star is redshifted to longer and longer wavelengths, and finally, according to theory, the star vanishes as an invisible black hole (Wiki NS). We will pursue black hole formation further in the next section.

10.4 Collapse of a Dust Star

For a sufficiently heavy stellar remnant internal pressure cannot halt the collapse to a black hole. In order to understand the process qualitatively we will make the drastic approximation of ignoring pressure entirely. The stellar model with no pressure is often called a *dust ball* or *dust star*. It will give us a rough qualitative picture of what happens in the collapse of a real star, and is an easy theoretical exercise. In fact we have already done all the mathematics needed and only some additional words are required.

Specifically, our model is a spherically symmetric ball of dust or gas with negligible internal pressure, which therefore collapses under the influence of gravity. This is illustrated in Fig. 10.5.

It is important to emphasize that we do not need to know the metric in the interior to understand the exterior, only that the exterior metric is Schwarzschild.

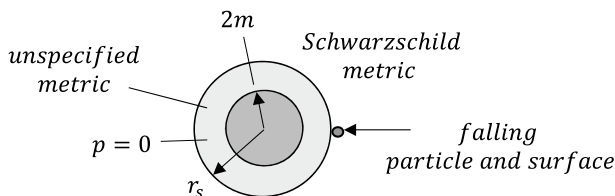


Fig. 10.5 The idealized model dust star. The exterior metric is Schwarzschild and the interior need not be specified

Consider a dust particle at or slightly inside the dust ball surface. It can make no difference in its behavior if we think of it as removed an arbitrarily small distance outside the star into the exterior space, since the only force acting on it is gravity. But we have already analyzed the fall of such a particle in the Schwarzschild spacetime in Sect. 10.1. Equations (10.12) and (10.13) and Fig. 10.2 summarize the results, that the particle, and thus the surface of the star, falls asymptotically to $2m$. The dust star collapses to a black hole. The collapse is quite rapid and effectively complete (see Exercise 10.7). The star is essentially *frozen* forever, and was thus originally called by Oppenheimer and Snyder a *frozen star* (Oppenheimer 1939).

Note carefully that the theoretical black hole, formed from the dust ball collapse, is full of matter out to the Schwarzschild radius for all time, as considered by an exterior observer using Schwarzschild time. Most important, the interior is *not empty* space.

In the above we did not explicitly need any properties of the interior of the dust ball to understand the surface and exterior behavior. However it is possible to model the entire dust ball collapse including the interior. Probably the easiest way to do this is to use for the interior a well-known metric from cosmology that describes one of the simplest models of the universe that we will discuss in a later section on cosmology; this was the seminal model developed by Oppenheimer and Snyder (Oppenheimer 1939). Since then there have been many variations of such analytic models of collapse, including nonuniform dust density and nonzero pressure (Adler 2005).

Many detailed and realistic models of collapse have also been studied using numerical methods and including rotation of the collapsing star. In general they verify the qualitative properties we have just discussed (Wiki GC).

10.5 Spinning Black Holes and the Kerr Metric

It is natural to expect that a non-rotating spherically symmetric star may collapse to form a spherically symmetric black hole as we have discussed. However we should not expect a spinning star to collapse into such a black hole since there is a preferred axis of rotation and angular momentum that must be conserved. It is now believed that a spinning star may collapse to form a different object, a spinning black hole, which

is the generalization of the Schwarzschild black hole. The solution for such an object was discovered by Kerr in 1963, many years after the Schwarzschild solution (Kerr 1963; Schiffer 1973; Adler 1975). The solution of the field equations is sufficiently lengthy that we will only give the final metric solution. In spherical coordinates it is given by the somewhat lengthy expression

$$\begin{aligned} ds^2 = & \left(1 - \frac{2mr}{r^2 + a^2 \cos^2 \theta}\right) c^2 dt^2 - \left(\frac{r^2 + a^2 \cos^2 \theta}{r^2 + a^2 - 2mr}\right) dr^2 \\ & - (r^2 + a^2 \cos^2 \theta) d\theta^2 - \left[(r^2 + a^2) \sin^2 \theta + \frac{2mra^2 \sin^4 \theta}{r^2 + a^2 \cos^2 \theta}\right] d\varphi^2 \\ & - 2\left(\frac{2mra \sin^2 \theta}{r^2 + a^2 \cos^2 \theta}\right) c dt d\varphi, \quad \text{Kerr metric.} \end{aligned} \quad (10.25)$$

Notice that the metric tensor components are independent of both t and φ and the solution is axially symmetric but not spherically symmetric. As with the Schwarzschild solution the geometric mass parameter m is related to the mass M of the source by $m = GM/c^2$ and has the dimension of a length. The other parameter a in the metric is related to the angular momentum J by $ma = -GJ/c^3$ and is also a length. We may refer to it as the geometric angular momentum.

The Kerr black hole or spinning black hole has an interesting horizon structure that is unlike the Schwarzschild black hole. There is an infinite redshift surface which we find by setting the metric term $g_{00} = 0$, giving

$$r_\infty = m + \sqrt{m^2 - a^2 \cos^2 \theta}. \quad (10.26)$$

This agrees with the Schwarzschild infinite redshift surface for $a = 0$. An emitting atom at this surface will have its radiation shifted to zero frequency at large radial distances, as in the Schwarzschild case.

It is also interesting to find the surface that is a null surface, or one-way membrane or horizon; this is the true black hole surface since no physical object may emerge from it. It is not hard to find the null surface using the results of Sect. 10.2; it is

$$r_{ns} = m + \sqrt{m^2 - a^2}. \quad (10.27)$$

Again this equals the Schwarzschild black hole surface when $a = 0$. Notice that the null or black hole surface is spherical and *inside* the infinite redshift surface, and the region between is an oblate shell. That region has some peculiar properties in that a body there may have negative total energy—that is its gravitational potential energy may exceed its rest energy. Note also that the angular momentum parameter a may not exceed the mass parameter m or the null surface and infinite redshift surface both become imaginary and meaningless.

Like the Schwarzschild metric the Kerr metric is believed to describe the exterior of a collapsed star, but not the interior. The interior problem is sufficiently complicated

that there is no known analog of the dust ball collapse model to describe the collapse of a spinning model star.

It is worth noting that if the black hole is slowly spinning, so that a is small, then the Kerr metric is well approximated by the first order expansion

$$ds^2 = \left(1 - \frac{2GM}{c^2 r}\right) c^2 dt^2 - \left(1 - \frac{2GM}{c^2 r}\right)^{-1} dr^2 - r^2 (d\theta^2 + \sin^2 \theta d\varphi^2) - \frac{4ma}{r} \sin^2 \theta c dt d\varphi \quad \text{Kerr metric to first order in } a, \quad (10.28)$$

which we recognize as the Schwarzschild metric plus a cross term. This metric was discovered by Lense and Thirring only a few years after the advent of general relativity (Thirring 1918). It is very simple and convenient for astrophysical applications since it describes the exterior of slowly spinning roughly spherical bodies rather well.

10.6 Black Holes in the Real Universe

A few brief comments are in order on actual black holes in nature. Theorists generally agree that a non-radiating condensed stellar-type object with a mass greater than the TOV limit cannot be a neutron star; by default it must be a black hole (Misner 1973). A number of high energy flickering X-ray sources are thus likely to be black holes. Such X-ray sources are generally thought to be black holes with material from a companion star falling into them; the material should spiral in an *accretion disk* into the large gravitational potential energy field of the black hole and emit X-rays as it is heated to very high temperatures. This is illustrated in Fig. 10.6; also see Exercise 10.12.

Historically the X-ray source Cygnus x-1 was the first widely accepted black hole; it was discovered in 1964. Its X-ray emissions flicker on a millisecond scale, indicating that the system is very small, less than c over the flicker frequency, of order 100 km. Its mass is about 15 solar masses. Since then many such black holes have been observed and are now a commonplace in astronomy.

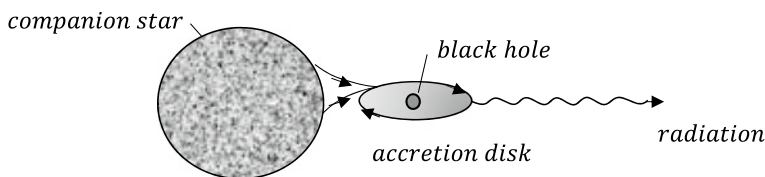


Fig. 10.6 Material from a companion star falls into a black hole, forming an accretion disk as it spirals in. Heating of the material by the gravitational energy produces diverse electromagnetic radiation such as X-rays

In Sects. 11.3–11.6 we will discuss gravitational waves. The source of the first waves detected was the merger of two black holes from close orbit; the 2016 event was called GW150914 and involved two black holes of about 30 solar masses each (Abbott 2016). Since then there have been many other similar gravitational wave events detected, including the merger of two neutron stars. It is rather remarkable that those detections involve two of the most extraordinary things predicted by general relativity, black holes and gravitational waves. Moreover they provide the first data we have obtained on truly strong gravitational fields and give some of the strongest evidence for the existence of black holes.

Black holes are not limited to stellar scale objects. There is no reason why the processes that give rise to clusters of stars and galaxies should not also produce black holes of much larger than stellar mass. For a star cluster the presence of a supermassive black hole (SMBH) in the center would be signaled by rapid motion of stars near the center; such large kinetic energy implies large potential energy, which in turn implies a small massive object at the center. Just such clusters have been observed. In particular the center of our Milky Way galaxy contains a very interesting black hole of about 4 million solar masses, called Sagittarius A* or Sgr A*: it is not visible to optical telescopes due to obscuring dust and gas but is the object of much present research activity using radio telescopes designed to measure the motion of stars near the central black hole. The goal is to study the system as close to the Schwarzschild radius as possible.

It is widely thought that many galaxies, perhaps most, contain a SMBH at their centers. There is a class of galaxies with “active galactic nuclei” (AGN) which emit intense radiation and fit this picture. It is probable that quasars, which emit enormous amounts of electromagnetic radiation, are powered by black holes at their centers. The mechanism is analogous to that in Fig. 10.6, but on a larger scale. A SMBH at the center of the galaxy Messier 87 has actually been imaged using a global network of radio telescopes, called the event horizon telescope (EHT), set up to act like an interferometer; the image is a rather fuzzy ring as shown in Fig. 10.7. A false-color image and more details are available on a number of websites (EHT 2019; Wiki BH).

In summary, black holes have become very well-known among physicists and astronomers and even the educated public. They occur at scales from stellar to galactic and are one of the prime focus areas of current research.

10.7 Hawking Radiation from a Black Hole

All of the preceding discussion of black holes was based on classical physics, and ignored quantum effects. Such quantum effects have been and still are the focus of much theoretical activity. We will discuss here only a simple version of the most notable effect, the thermal radiation emitted by a black hole, called Hawking radiation. Before we begin we emphasize that quantum effects such as Hawking radiation have *not* been observed despite strong efforts and remain in the realm of theoretical speculation.



Fig. 10.7 The supermassive black hole in Messier 87 as imaged by the EHT

There are two elementary ways to heuristically obtain the Hawking formula for the temperature of a black hole, one using the uncertainty principle, and one using the second law of thermodynamics. The quantum field theory derivation used originally by Hawking is beyond our present scope so we will use the uncertainty principle derivation even though it is heuristic and crude (Hawking 1974; Adler 2001, 2006). The derivation based on thermodynamics is discussed in Exercise 10.15 (Ohanian 1994).

Our derivation uses the uncertainty principle combined with some qualitative concepts from quantum field theory, so it is better motivated and more convincing than dimensional analysis alone. In field theory we find that the vacuum is not at all empty but is filled with *virtual* particles interacting as symbolized by Feynman diagrams, one of which is shown in Fig. 10.8: an electron and positron and photon materialize out of nothing and have a fleeting existence before they recombine and vanish.

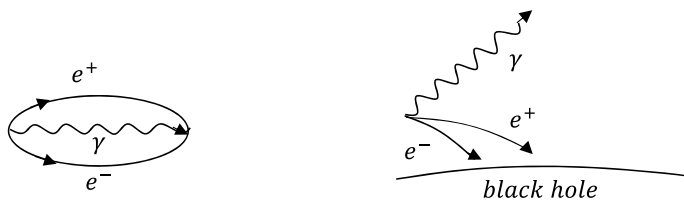


Fig. 10.8 On the left is a vacuum bubble diagram of quantum electrodynamics. The particles have only a brief existence before they are forced by energy conservation to recombine and vanish. On the right the photon may escape since the black hole can provide energy to the system

The brief violation of energy conservation is allowed according to the energy-time version of the uncertainty principle; this relates the lifetime of a state to the uncertainty in its energy by

$$\Delta E \Delta t \approx \hbar. \quad (10.29)$$

Thus the particles in the left Fig. 10.8 with energy ΔE equal to the combined rest mass or greater can only exist for a time Δt . However, near a black hole we can imagine that the electron and positron fall toward the black hole rather than recombining with the photon, while the photon escapes outward as in the right of Fig. 10.8. Rather than being forced to recombine by energy conservation the photon becomes real and escapes with energy provided by the black hole.

Consider the escaping particle and the uncertainty principle in its usual form

$$\Delta p \Delta x \approx \hbar/2. \quad (10.30)$$

This lets us estimate the momentum and energy the escaping photon can have. The only scale in the system is the Schwarzschild radius of the black hole $r_s = 2m$, so we take that to be the uncertainty Δx and have for the photon momentum

$$p \approx \Delta p \approx \hbar/2\Delta x \approx \hbar/4m, \quad (10.31)$$

(See Exercise 10.13 for a stronger motivation for this choice). This gives for the photon energy

$$E = pc \approx \frac{\hbar c}{4m} = \frac{\hbar c^3}{4GM}. \quad (10.32)$$

If we now assume that the spectrum of such escaping photons is thermal then the temperature and energy are related by

$$kT \approx E \approx \frac{\hbar c^3}{4GM}. \quad (10.33)$$

This is about the same relation for the temperature obtained by Hawking using quantum field theory, except that his result contains 8π rather than 4 in the denominator,

$$kT_H = \frac{\hbar c^3}{8\pi GM}, \quad \text{Hawking temperature.} \quad (10.34)$$

The result (10.33) can also be obtained with a thermodynamic argument, as outlined in Exercise 10.15 (Ohanian 1994).

Having obtained the temperature of a black hole we can also calculate an entropy. We imagine building the black hole by assembling small masses, each with rest

energy $dQ = dMc^2$. According to the thermodynamic definition of entropy S the increase for each such small mass is

$$dS = \frac{dQ}{kT} = \frac{8\pi GM}{\hbar c} dM. \quad (10.35)$$

(Note that we here use a dimensionless definition of entropy.) The total entropy of the black hole is thus

$$S = \frac{4\pi G}{\hbar c} M^2 = \frac{1}{4} \frac{A_{\text{BH}}}{\ell_{\text{P}}^2}, \quad \text{where } A_{\text{BH}} = 4\pi r_s^2, \\ \ell_{\text{P}} = \sqrt{\frac{c^3}{G\hbar}} = 1.6 \times 10^{-35} \text{ m}. \quad (10.36)$$

It is equal to one fourth of the area of the black hole A_{BH} divided by the square of the Planck distance ℓ_{P} ; the Planck distance may in fact be the smallest physically meaningful distance as we will discuss in a later section on the Planck scale. The result (10.36) is called the Bekenstein entropy and was obtained, up to a factor, by Bekenstein before Hawking obtained his result (Bekenstein 1973). The Bekenstein formula simply assigns one unit of entropy to each tiny Planck scale area $4\ell_{\text{P}}^2$ on the black hole surface.

The black hole entropy has the peculiar feature that it is proportional to a 2-dimensional surface area. It is more common that entropy, called an extensive property, is proportional to the volume of a system. Some theorists have made much of this fact and have introduced the “holographic principle,” that the information on the surface of the black hole is somehow equivalent to the information one expects for the volume of the black hole; some have tried to elevate this to a general principle. This is of course highly speculative.

Hawking’s formula for the temperature is quite remarkable since it predicts a specific phenomenon that involves both gravity and quantum mechanics. Hawking radiation has not yet been observed in the real world. For stellar mass black holes we do not expect to see it since the temperature predicted by (10.34) is only about 6×10^{-8} K. This is much less than the ambient 2.7 K cosmic background radiation that permeates the universe, so a black hole at such a low temperature would absorb more radiation than it emits. For sufficiently small black holes, with temperatures greater than the background radiation we should be able to detect the radiation; the relevant mass is roughly $10^{-8} M_{\odot}$. Indeed, as such a small black hole radiates it will lose energy and thus become lighter, so its temperature will increase without limit according to (10.34), and it should end up emitting a very bright flash at the end of its life. See Exercise 10.17. The absence of observations of such flashes could be due to the lack of small black holes or the incorrectness of the theory.

The late stages of black hole evaporation likely involve very large energies and small distances, of order the Planck scale of 10^{19} GeV and 10^{-35} m. We of course have no experimental knowledge of such things, and thus no dependable theory, so the end product of the evaporation is unknown. It is widely believed that distances

smaller than the Planck distance are not physically meaningful, so it is possible that black holes might not evaporate entirely away and could leave a remnant of order the Planck size. Such remnants could be a candidate for the dark matter particles that will be discussed in Chap. 16; we will come back to this also in Appendix 3 in Chap. 19 (Adler 2001).

Exercises

- 10.1 Calculate the geometric mass m of the sun in km. What is the radius of the sun in km? How much does the Schwarzschild metric component $g_{00} = 1 - 2m/r$ differ from 1 near the surface of the sun? Repeat the calculations for the earth. Do you see why it is difficult to do interesting general relativity experiments in the solar system and especially on the surface of the Earth?
- 10.2 What is the area of the black hole surface? This area plays an interesting role in the quantum mechanical properties of black holes, as we have discussed in Sect. 10.7 on Hawking radiation.
- 10.3 Evaluate the Riemann curvature tensor to see if it is singular or zero at the Schwarzschild radius. (You may instead simply look up the Riemann tensor for the Schwarzschild metric.)
- 10.4 Is the Ricci tensor singular at the Schwarzschild radius? Is the Riemann scalar singular at the Schwarzschild radius? Hint: this requires no calculation.
- 10.5 Verify that a photon falling onto a black hole approaches it exponentially; that is verify (10.6).
- 10.6 Verify that a particle falling onto a black hole approaches it exponentially; that is verify (10.13).
- 10.7 Consider a particle (or photon) falling onto a black hole surface. Take the geometric mass to be about 5 km and the initial position to be about 5 km. Calculate approximately how far away the particle is after, 10^{-9} , 10^{-5} , 1 s, 1 year. Does it really make *physical sense* to say that the particle never quite reaches the surface, or that a collapsing dust star never quite becomes a black hole? See Sect. 19.6 for comments on small distances.
- 10.8 In Sect. 10.1 of the text we refer to an outside observer situated far from the black hole,. Repeat the discussion of the red shift using an observer at a finite fixed distance outside the black hole using laboratory time intervals $t_{ob} = \sqrt{1 - 2m/r_{ob}} t$. Do any of the important qualitative conclusion concerning the fall of a particle to the surface change significantly?
- 10.9 Using some reference on the Kruskal Szekeres coordinates, such as Adler (1975), show how the region inside an empty theoretical black hole surface is only relevant to outside observers for $t > \infty$; that is we in the exterior simply cannot communicate with that interior region.
- 10.10 Make a qualitative sketch of the trajectories of light traveling radially inward and radially outward in the exterior of a black hole with Schwarzschild geometry. In the sketch draw the local light cones for such radial motion, and notice that they degenerate at the Schwarzschild radius and lie along the surface.

- 10.11 Let's do some rough order of magnitude energy conversion estimates. Using a convenient reference estimate what fraction of the rest energy is liberated in a typical atomic or molecular reaction? What of a nuclear reaction? If a mass falls onto the surface of a neutron star estimate roughly what fraction of its rest energy turns into kinetic energy and then into heat. What of a mass falling into a black hole? (You may use a rough estimate for the potential energy of a particle in the Schwarzschild metric using the classical potential.)
- 10.12 There is a heuristic motivation for taking the uncertainty in the position of a particle near a black hole to be the Schwarzschild radius $2m$. One may calculate the electric field of a small electric charge near the black hole surface in the Schwarzschild metric; it turns out that the field lines wrap around the surface in such a way that at a large distance they appear to diverge from the center of the hole rather than a point near the surface where the charge actually is; this may be interpreted as an uncertainty in position. Study this using the references Ruffini (1971) and Adler (1976, 2001).
- 10.13 Can you think of a heuristic way to show that the hawking radiation is specifically thermal? (Nobody else has done this!)
- 10.14 The second heuristic way to estimate the Hawking temperature is to do a heat engine gedanken (thought) experiment. It goes like this: fill a box with thermal radiation from a hot heat reservoir far from a black hole; lower the box to the surface of the black hole to run an engine and do work; at the surface, taken to be a cold reservoir, release the radiation to the surface; pull the empty box back to the hot reservoir and repeat. The ideal efficiency of such a heat engine, using the second law of thermodynamics, gives a rough estimate for the effective temperature of the black hole. Note that the necessary minimum size of the box is important. See Ohanian (1994).
- 10.15 Calculate the Hawking temperature for a black hole of solar mass, as given in the text.
- 10.16 Use the Stefan-Boltzmann law of radiation for a black body to calculate the energy radiated by a black hole. Use that result to estimate the lifetime of a black hole before it completely evaporates away. See also Chap. 19 on the end stages of black hole evaporation.
- 10.17 Ponder for yourself the idea that the black hole entropy appears to reside on a 2-surface; does it really violate any physical laws or intuition? It may be interesting to read some of the papers on the holographic principle. See also the comments on the Planck scale in Chap. 19.

Chapter 11

Linearized General Relativity and Gravitational Waves



Abstract Gravitational waves are the analog of radio waves in electromagnetic theory. They were first predicted soon after the advent of general relativity theory, and after about a century of theoretical research and decades of experimental work they have been finally detected. In this chapter we develop the theory of the production, the propagation, and the detection of gravitational waves. Gravitational waves provide an entirely new observational window on the universe; the mergers of black holes and neutron stars are the sources of the waves so far observed.

11.1 The Field Equations of the Linearized Theory

In this chapter we will discuss approximate solutions to the Einstein equations, with an emphasis on gravitational waves. Einstein recognized when he first formulated his equations that exact solutions would be difficult to obtain since the equations are nonlinear. To get approximate solutions to the field equations we will linearize them by assuming, as we did in Chaps. 7 and 8, that the metric is the Lorentz metric plus a small dimensionless perturbation that describes weak gravity. That is

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}, \text{ all } h_{\mu\nu} \ll 1. \quad (11.1)$$

Then the inverse metric is given to lowest order by

$$g^{\mu\nu} = \eta^{\mu\nu} - h^{\mu\nu}, \quad h^{\mu\nu} \equiv \eta^{\mu\alpha} h_{\alpha\beta} \eta^{\beta\nu}. \quad (11.2)$$

We will usually call the perturbation $h_{\mu\nu}$ the *metric field*. In this section we will often not bother to repeat the phrases “approximately equal” or “to lowest order in the perturbation” but assume that (almost) all of the equations are approximate and only correct to lowest order. Note moreover that indices may usually be raised and lowered with the Lorentz metric, appropriate to this approximation, as in (11.2). Since the Lorentz metric has constant elements many manipulations are thereby greatly simplified. Many of the algebraic manipulations are the same as in special relativity.

The connections are, from the definition (5.19),

$$\Gamma_{\beta\gamma}^{\alpha} = \frac{1}{2}\eta^{\alpha\lambda}(h_{\beta\lambda,\gamma} + h_{\gamma\lambda,\beta} - h_{\beta\gamma,\lambda}). \quad (11.3)$$

The approximate Riemann tensor has only two terms,

$$R^{\lambda}_{\beta\gamma\delta} = \Gamma_{\beta\gamma,\delta}^{\lambda} - \Gamma_{\beta\delta,\gamma}^{\lambda}. \quad (11.4)$$

Using the symmetry of the metric and the commutativity of ordinary derivatives we can then calculate the Riemann tensor to be

$$\begin{aligned} R^{\lambda}_{\beta\gamma\delta} &= \frac{1}{2}\eta^{\lambda\tau}(h_{\gamma\tau,\beta} + h_{\beta\tau,\gamma} - h_{\beta\gamma,\tau})_{,\delta} - \frac{1}{2}\eta^{\lambda\tau}(h_{\delta\tau,\beta} + h_{\beta\tau,\delta} - h_{\beta\delta,\tau})_{,\gamma} \\ &= \frac{1}{2}\eta^{\lambda\tau}(h_{\gamma\tau,\beta,\delta} - h_{\beta\gamma,\tau,\delta} - h_{\delta\tau,\beta,\gamma} + h_{\beta\delta,\tau,\gamma}). \end{aligned} \quad (11.5)$$

The fully covariant form is

$$R_{\alpha\beta\gamma\delta} = \frac{1}{2}(h_{\gamma\alpha,\beta,\delta} - h_{\beta\gamma,\alpha,\delta} - h_{\delta\alpha,\beta,\gamma} + h_{\beta\delta,\alpha,\gamma}). \quad (11.6)$$

The last three expressions of course are consistent with (8.15), (8.16) and (8.17) for the Riemann tensor in the geodesic system.

The Reimann tensor (11.6) has a remarkable property under what we may call a “small” change of coordinates; the coordinate change is a close analog of a gauge transformation in electromagnetism. By a small change we mean a transformation to a primed system using four arbitrary functions f^{α} , of the form

$$\begin{aligned} x'^{\alpha} &= x^{\alpha} - f^{\alpha}(x^{\beta}), \\ \frac{\partial x'^{\alpha}}{\partial x^{\sigma}} &= \delta_{\sigma}^{\alpha} - f^{\alpha}_{,\sigma}, \quad \frac{\partial x^{\gamma}}{\partial x'^{\rho}} = \delta_{\rho}^{\gamma} + f^{\gamma}_{,\rho}, \quad f^{\alpha}_{,\sigma} \ll 1, \quad \text{small.} \end{aligned} \quad (11.7)$$

This coordinate transformation we may also call a *gauge* change or *gauge* transformation. It is easy to see that the metric remains nearly Lorentzian under such a change, with

$$h'_{\mu\nu} = h_{\mu\nu} + (f_{\mu,\nu} + f_{\nu,\mu}). \quad (11.8)$$

What is remarkable and interesting is that if we calculate the Riemann tensor for this metric we see that it is composed of two parts, one for each term on the right side of (11.8); the second part, that which depends on $f_{\nu,\mu}$, is identically zero. From this it follows that the Riemann tensor is invariant under the gauge transformation,

$$R'_{\alpha\beta\gamma\delta} = R_{\alpha\beta\gamma\delta}. \quad (11.9)$$

The situation is in close analogy to gauge invariance in electromagnetism: a gauge change of the vector potential does not change the electromagnetic fields; the gauge change of the coordinates does not change the Reimann tensor, which is the intrinsic indicator of the gravitational field as we discussed in Chap. 8.

The expression $f_{\mu,v} + f_{v,\mu}$ that appears in (11.8) gives a null Riemann tensor or flat space, and thus must be a solution of the field equations; it is called a Weyl solution. We will find the idea of gauge transformations and the gauge invariance of the Reimann tensor very useful when we study gravitational waves.

The Ricci tensor and Riemann (or Ricci) scalar follow from contraction of the Reimann tensor (11.6). Using the symmetry of the metric and second derivatives and raising and lowering indices with the Lorentz metric we find for the Ricci tensor and the Riemann scalar,

$$R_{\mu\nu} = \frac{1}{2}(h_{,\mu,v} + h_{\mu\nu,\lambda}{}^{,\lambda} - h^\lambda{}_{v,\lambda,\mu} - h^\lambda{}_{\mu,\lambda,v}), \quad h \equiv h^\alpha{}_\alpha = \eta^{\beta\alpha} h_{\alpha\beta}, \quad (11.10a)$$

$$R = h^\alpha{}_{,\alpha} - h^{\alpha\omega}{}_{,\alpha,\omega}. \quad (11.10b)$$

As usual the upper indices are raised with the Lorentz metric, including the derivatives.

From these the Einstein tensor follows easily. Setting the Einstein tensor equal to the energy-momentum tensor then gives the linearized field equations,

$$\begin{aligned} G_{\mu\nu} &= \frac{1}{2}[(h_{,\mu,v} + h_{\mu\nu,\lambda}{}^{,\lambda} - h^\lambda{}_{v,\lambda,\mu} - h^\lambda{}_{\mu,\lambda,v}) - \eta_{\mu\nu}(h^\alpha{}_{,\alpha} - h^{\alpha\omega}{}_{,\alpha,\omega})] \\ &= -\left(\frac{8\pi G}{c^2}\right)T_{\mu\nu}. \end{aligned} \quad (11.11)$$

Note that here the energy momentum tensor here has units of mass density.

It is possible and quite useful to simplify the field equations before we attempt to solve them. In order to eliminate the terms containing the trace h in (11.11) we define an object with a modified trace and write it with a hat,

$$\bar{h}_{\alpha\beta} \equiv h_{\alpha\beta} - \frac{1}{2}\eta_{\alpha\beta}h, \quad \text{so} \quad \bar{h} = -h, \quad h_{\alpha\beta} = \bar{h}_{\alpha\beta} - \frac{1}{2}\eta_{\alpha\beta}\bar{h}. \quad (11.12)$$

The second and third equations in (11.12) follow easily from the first. (To avoid confusion, we will never in this chapter use a hat to indicate a transformed coordinate system. Note also that the new object should be referred to as $\hat{h}$ and never $\bar{h}$, which name is reserved for the reduced Planck constant.) In terms of the $\bar{h}_{\alpha\beta}$ the linearized field equations simplify to

$$\begin{aligned}
G_{\mu\nu} &= \frac{1}{2} \left[\left(\bar{h}_{\mu\nu}{}^{,\lambda}{}_{,\lambda} - \bar{h}_\nu{}^{\lambda}{}_{,\lambda,\mu} - \bar{h}_\mu{}^{\lambda}{}_{,\lambda,\nu} \right) + \eta_{\mu\nu} (\bar{h}^{\alpha\omega}{}_{,\omega,\alpha}) \right] \\
&= - \left(\frac{8\pi G}{c^2} \right) T_{\mu\nu}.
\end{aligned} \tag{11.13}$$

The traces have disappeared as desired. We have also rearranged the dummy indices to be more suggestive for the next simplification.

There is a gauge or coordinate transformation that will make the bracket in (11.13) yet simpler by eliminating three of the terms that contain the divergence $\bar{h}^{\alpha\lambda}{}_{,\lambda}$ leaving only one term on the left. From (11.8) and (11.12) we may calculate the transformation of $\bar{h}^{\mu\nu}$ to a new primed coordinate system

$$\bar{h}'^{\mu\nu} = \bar{h}^{\mu\nu} + (f^{\mu,\nu} + f^{\nu,\mu}) - \eta^{\mu\nu} f^{\lambda}{}_{,\lambda}. \tag{11.14}$$

Then its divergence in the primed system is

$$\bar{h}'^{\mu\nu}{}_{,\nu} = \bar{h}^{\mu\nu}{}_{,\nu} + f^{\mu,\nu}{}_{,\nu}. \tag{11.15}$$

If we desire this divergence to be conveniently zero in the primed system we need only choose the functions f^μ to satisfy a Poisson equation

$$f^{\mu,\nu}{}_{,\nu} = -\bar{h}^{\mu\nu}{}_{,\nu}. \tag{11.16}$$

In general there exists a solution for this equation, so the divergence terms in the field equations vanish in the primed system and we are left with a rather simple set of field equations in that system

$$G_{\mu\nu} = \frac{1}{2} \bar{h}_{\mu\nu}{}^{,\lambda}{}_{,\lambda} = - \left(\frac{8\pi G}{c^2} \right) T_{\mu\nu}, \tag{11.17a}$$

$$\bar{h}^{\mu\nu}{}_{,\nu} = 0. \tag{11.17b}$$

We consider forthwith only systems where the divergence is zero and thus have dropped the primes in (11.17). This may also be written using the d'Alembertian operator $\square^2$ as

$$\bar{h}_{\mu\nu}{}^{,\lambda}{}_{,\lambda} = \square^2 \bar{h}_{\mu\nu} = - \left(\frac{16\pi G}{c^2} \right) T_{\mu\nu}, \tag{11.18a}$$

$$\bar{h}^{\mu\nu}{}_{,\nu} = 0, \quad \square^2 \equiv \eta^{\alpha\beta} \frac{\partial}{\partial x^\alpha} \frac{\partial}{\partial x^\beta} = \frac{\partial^2}{\partial t^2} - \nabla^2. \tag{11.18b}$$

This is a somewhat nonstandard notation for the d'Alembertian, using a square, in analogy with the Laplacian ∇^2 in three dimensions. As we noted above, in the last two lines we have dropped the prime notation, assuming that we always work in a system where the divergence is zero. Notice that the equations in (11.17) and (11.18) are consistent since the energy-momentum tensor has a zero divergence as we noted previously.

The gauge choice we used in the above paragraph is often called the *Lorentz gauge* since it is the exact analog of the Lorentz gauge of electromagnetism, but it is also often called the de Donder or harmonic gauge. Its utility is obvious because of the way it simplifies the field equations to be a set of Poisson equations with a divergence constraint. They could hardly be more simple. Unless noted otherwise we will always work in the Lorentz gauge. In the final form (11.18) the field equations are very similar to many in classical physics, in particular electromagnetic radiation theory, so many problems in gravity may be solved using well-known techniques, as we discuss in the Appendices.

There is one more important fact to be obtained from (11.15) for the transformation of the metric field divergence. If the original gauge is Lorentz and we transform with functions f^α to a new gauge then the new gauge is also Lorentzian if and only if the f^α obey the wave equation $\square^2 f^\alpha = 0$. We will use this sort of gauge transformation to great advantage in what follows.

11.2 The Classical Limit

By the classical limit we mean that the gravitational fields are weak and independent of time, and the source is the matter density, independent of velocity as in Poisson's equation. We have already treated this situation in Chaps. 7 and 8 where we developed basic ideas and field equations, but in this section we will go a little deeper and be more general and systematic.

The source for the classical limit case must be well-described by the energy-momentum tensor of slowly moving dust, as we discussed in Chap. 8, with only a 0,0 metric component since the motion is to be neglected. That is

$$T_{\mu\nu} = \begin{pmatrix} \rho & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \quad (11.19)$$

For a time independent system the field equations (11.18) then give

$$\nabla^2 \bar{h}_{00} = -\left(\frac{16\pi G}{c^2}\right)\rho. \quad (11.20)$$

This is the same as Poisson's equation for the classical potential ϕ , so we identify a relation between the 0,0 metric component and the classical potential

$$\bar{h}_{00} = \frac{4}{c^2}\phi. \quad (11.21)$$

The other components of $\bar{h}_{\alpha\beta}$ we may take to be zero, which is consistent with the energy-momentum tensor (11.19). To get the physical metric field $h_{\mu\nu}$ we use its relation to $\bar{h}_{\mu\nu}$ in (11.12) and obtain the metric field and line element in terms of ϕ

$$h_{\mu\nu} = \frac{2}{c^2}\phi\delta_{\mu\nu}, \quad ds^2 = \left(1 + \frac{2\phi}{c^2}\right)c^2dt^2 - \left(1 - \frac{2\phi}{c^2}\right)d\vec{x}^2. \quad (11.22)$$

This constitutes a rather general solution for the field created by low density matter moving slowly, giving the metric in terms of the classical potential. In particular for a stationary point mass M we have a line element

$$ds^2 = \left(1 - \frac{2GM}{c^2r}\right)c^2dt^2 - \left(1 + \frac{2GM}{c^2r}\right)d\vec{x}^2, \quad r = \sqrt{\vec{x}^2}. \quad (11.23)$$

Equations (11.22) and (11.23) are useful in many practical applications in celestial mechanics.

Note that (11.23) is spatially isotropic and thus does not have the same form as the Schwarzschild solution in Schwarzschild coordinates (9.19); instead it has the same form as the Schwarzschild solution in isotropic coordinates (9.59) and the Eddington form (9.60), which is useful in discussions of the experimental tests of relativity (Will 2014).

11.3 Gravitational Plane Waves

One of the most interesting properties of the linearized field equations is that they admit wave solutions, much as Maxwell's electrodynamics admits electromagnetic wave solutions. Most of what we do in this section is very similar to the solutions of Maxwell's equations for electromagnetic waves in terms of the 4-vector potential and the Maxwell tensor, so the reader who is familiar with that topic will find it especially easy. For those not familiar with electromagnetic waves or who desire a brief review see Appendices 2 and 3.

In vacuum the field equations (11.18) are

$$\bar{h}_{\mu\nu}{}^{,\lambda}{}_{,\lambda} = \square^2 \bar{h}_{\mu\nu} = 0, \quad \bar{h}^{\mu\nu}{}_{,\nu} = 0. \quad (11.24)$$

The first of these is the wave equation with velocity equal to c and the second is the Lorentz gauge condition. From (11.24) it is apparent that all of the functions $\bar{h}_{\mu\nu}$ and $h_{\mu\nu}$ and h and $\bar{h}$ obey the wave equation. Moreover it is important that the Riemann tensor does also, as is clear from (11.6).

To get a solution for a plane wave moving at c in the spacetime direction k_μ we choose a smooth function $\ell(U)$ of the scalar $U = k_\beta x^\beta$. Then its derivatives obey,

$$\ell(U) = \ell(k_\beta x^\beta), \quad \frac{\partial}{\partial x^\mu} \ell = \frac{d\ell}{dU} \frac{\partial U}{\partial x^\mu} = \ell_{,U} k_\mu, \quad \square^2 \ell = \ell_{,U,U} k_\mu k^\mu. \quad (11.25)$$

Here the comma U indicates a derivative with respect to the scalar argument U . Any such function is thus a solution of the wave equation if k_μ is chosen to be a null vector, $k_\mu k^\mu = 0$. We may choose the z direction to lie along the space component of the null vector k_μ so it is a constant multiple of $(1, 0, 0, -1)$ and U is a multiple of $ct - z$. We may thus take $U = ct - z$ without loss of generality in what follows.

To set up a plane wave solution we write the metric field in terms of the function $\ell(U)$ and a set of coefficients $\epsilon_{\mu\nu}$,

$$\bar{h}_{\mu\nu}(U) = \epsilon_{\mu\nu} \ell(U). \quad (11.26)$$

The $\epsilon_{\mu\nu}$ is a constant array of what we will call *polarization* coefficients, in analogy with electromagnetic radiation. It is only necessary to impose the Lorentz condition in (11.24) to determine the coefficients. To do this we align the z axis along the direction of the wave as above, then choose the coefficient matrix to be either of the following

$$\epsilon_{\mu\nu} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \text{or} \quad \epsilon_{\mu\nu} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{so} \quad \epsilon_{\mu\nu} k^\nu = 0. \quad (11.27)$$

The solution for the metric is then

$$h_{\mu\nu} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & h_{11} & h_{12} & 0 \\ 0 & h_{12} & -h_{11} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \quad (11.28)$$

Here h_{11} and h_{12} are arbitrary smooth functions of $U = ct - z$. Note that the trace of the solution is zero, so $\bar{h}_{\mu\nu} = h_{\mu\nu}$. The gauge used in (11.28) is called the traceless transverse or TT gauge since $h_{\mu\nu}$ is traceless and only has components in the x, y plane, perpendicular to the direction of propagation. The line element is

$$ds^2 = c^2 dt^2 - (1 - h_{11})dx^2 - (1 + h_{11})dy^2 + 2h_{12}dxdy - dz^2. \quad (11.29)$$

In the next section we will study the meaning of the two arbitrary functions in the metric.

A comment on some basic physics is in order at this point: the metric obeys the wave equation with velocity c so one might think that this shows the physical gravitational field propagates at c . However, this is not the correct viewpoint since the Riemann tensor is the intrinsic signature of gravity rather than the metric. But since the Riemann tensor also obeys the wave equation with velocity c we can indeed correctly say that the gravitational field propagates at c , at least in the weak field approximation.

We have easily obtained the above as one convenient plane wave solution for the metric field. It is important to also show that if one has any solution to the basic (11.24) then it can be put into the form (11.28) by a gauge transformation, and most important the gauge transformation can be obtained explicitly. This is very important for the solutions we will obtain in the section below on sources of gravitational waves. The remainder of this section will be devoted to showing in considerable algebraic detail how to transform an arbitrary plane wave solution to the traceless transverse form in (11.28) that serves as a canonical form.

The elements of the metric field as well as the gauge transformation functions may all be taken to be functions of $U = ct - z$, as we have discussed above. The Lorentz condition in (11.24) then strongly restricts the components of the metric, giving

$$\bar{h}^{\mu\nu}_{,v} = \bar{h}^{\mu 0}_{,0} + \bar{h}^{\mu 3}_{,3} = \bar{h}^{\mu 0}_{,U} - \bar{h}^{\mu 3}_{,U} = (\bar{h}^{\mu 0} - \bar{h}^{\mu 3})_{,U} = 0. \quad (11.30)$$

Since the components $\bar{h}^{\mu 0}$ and $\bar{h}^{\mu 3}$ are functions of only the variable U they can only differ by a constant; we will consider them equal since we are interested in time varying wave fields rather than constant fields, and thus find

$$\bar{h}^{\mu 0} = \bar{h}^{\mu 3}. \quad (11.31)$$

This restricts the hat metric field so that it has only 6 independent components, as displayed here,

$$\bar{h}^{\mu\nu} = \begin{pmatrix} \bar{h}^{00} & \bar{h}^{01} & \bar{h}^{02} & \bar{h}^{00} \\ \bar{h}^{01} & \bar{h}^{11} & \bar{h}^{12} & \bar{h}^{01} \\ \bar{h}^{02} & \bar{h}^{12} & \bar{h}^{22} & \bar{h}^{02} \\ \bar{h}^{00} & \bar{h}^{01} & \bar{h}^{02} & \bar{h}^{00} \end{pmatrix}. \quad (11.32)$$

Since we remain always in a Lorentz gauge this holds in all systems we use.

If we now use the transformation relation for the hat metric field in (11.14) we can make all the elements with $\mu = 0$ index equal to zero in some new primed system. We do this beginning with the 0,1 element, and recall that f_μ does not depend on x , but only on z and ct , so that

$$\bar{h}'_{01} = \bar{h}_{01} + f_{0,1} + f_{1,0} = \bar{h}_{01} + f_{1,0} = \bar{h}_{01} + f_{1,U}. \quad (11.33)$$

This component can thus be made zero by choosing

$$f_{1,U} = -\bar{h}_{01}, \quad f_1(U) = -\int^U \bar{h}_{01} dU'. \quad (11.34)$$

The same procedure makes the 0,2 element equal to zero in an obvious way. For the 0,0 element we obtain in similar fashion

$$\bar{h}'_{00} = \bar{h}_{00} + 2f_{0,0} - f^\beta{}_{,\beta} = \bar{h}_{00} + (f_{0,U} - f_{3,U}), \quad (11.35)$$

so the 0,0 element can be made zero by choosing

$$f_{0,U} - f_{3,U} = -\bar{h}_{00}, \quad f_0(U) - f_3(U) = -\int^U \bar{h}_{00} dU'. \quad (11.36)$$

Thus, in the primed system the metric field has been reduced to an array with nonzero elements in only the 1,2 block. Three equations for the four transformation functions have been determined in the process.

Finally, to determine the 1,2 block in the primed frame we use the transformation (11.14) again to find

$$\bar{h}'_{11} = \bar{h}_{11} + (f_{0,U} + f_{3,U}), \quad \bar{h}'_{22} = \bar{h}_{22} + (f_{0,U} + f_{3,U}), \quad \bar{h}'_{12} = \bar{h}_{12} \quad (11.37)$$

These two relations allow us to make the trace in the primed system zero. From (11.37) the new trace is

$$\bar{h}'_{11} + \bar{h}'_{22} = \bar{h}_{11} + \bar{h}_{22} + 2(f_{0,U} + f_{3,U}). \quad (11.38)$$

This will be zero if we choose

$$f_{0,U} + f_{3,U} = -\frac{1}{2}(\bar{h}_{11} + \bar{h}_{22}), \quad f_0 + f_3 = -\frac{1}{2} \int^U (\bar{h}_{11} + \bar{h}_{22}) dU'. \quad (11.39)$$

Most important, the new 1,2 block is, from (11.37) and (11.39),

$$\bar{h}'_{11} = \frac{1}{2}(\bar{h}_{11} - \bar{h}_{22}), \quad \bar{h}'_{22} = -\frac{1}{2}(\bar{h}_{11} - \bar{h}_{22}), \quad \bar{h}'_{12} = \bar{h}_{12}. \quad (11.40)$$

This completes the transformation to a traceless transverse canonical form as promised.

In summary of this section we may take the plane wave metric field to have the form in (11.28) by a gauge choice. Recall moreover that since the trace is zero the hat notation is not needed. The last (11.40) will also be useful in the section on gravitational wave sources, in which the metric field from a source does not automatically have the canonical traceless transverse form.

11.4 Motion of Test Bodies in Gravitational Waves

Our results for the metric field of gravitational waves in the previous section are only part of the story. We also need to know how bodies move under the influence of the waves to fully understand the physics. For this we will first work out the geodesic equations of motion for test bodies in the metric (11.29). This is most easily done using the procedure discussed in Chap. 5: recall that we define a Lagrangian with the same mathematical form as the line element but with differentials replaced by derivatives with respect to the line element, and from that obtain the Euler-Lagrange equations as the equations of motion.

Before we begin we must emphasize that bodies that are acted on by forces other than gravity are not in free fall and do not move on geodesics. For example interatomic forces much stronger than gravity act on the atoms in a meter stick and make it act nearly like a rigid body in that its length is very nearly constant. Obviously gravitational waves have very little effect on such bodies. See the comments below on the Newtonian equivalent force and Exercise 9.3. To an extremely good approximation we may assume that meter stick distances are not significantly affected by gravitational waves; but bodies in free fall react significantly to the waves.

Let us first look at the case of $h_{12} = 0$, that is a diagonal metric. According to our recipe the Lagrangian is obtained from the line element

$$L = c^2 \dot{t}^2 - (1 - h_{11})\dot{x}^2 - (1 + h_{11})\dot{y}^2 - \dot{z}^2, \quad h_{11} = h_{11}(z - ct). \quad (11.41)$$

The geodesic equations are the Euler-Lagrange equations of this Lagrangian, and are simply obtained as

$$\begin{aligned} \dot{x}(1 - h_{11}) &= \text{const.}, \quad \dot{y}(1 + h_{11}) = \text{const.}, \\ \ddot{z} + \frac{1}{2}h'_{11}(\dot{x}^2 - \dot{y}^2) &= 0, \quad c^2\ddot{t} - \frac{1}{2}h'_{11}(\dot{x}^2 - \dot{y}^2) = 0. \end{aligned} \quad (11.42)$$

Here the prime denotes a derivative with respect to the argument of h_{11} , or $U = ct - z$. The solution to these equations is surprisingly easy for bodies that are at rest initially, before the wave arrives. Both constants in the x and y equations are then equal to zero so $\dot{x} = \dot{y} = 0$ and bodies remain at the same x and y positions. This implies furthermore that from the z equation $\ddot{z} = 0$, so bodies initially at rest at $z = 0$ do not move in the z direction. Finally the t equation tells us that $c\dot{t} = 0$ so we may choose the coordinate time along the geodesic to be equal to the proper time, or $ct = s$.

In summary, the motion is very simple: bodies initially at coordinate rest in the traceless transverse gauge remain at coordinate rest as the wave passes. For this reason the coordinates are called *co-moving*. However coordinate distances and physical distances are different, so bodies at rest in the coordinate system are not physically at rest.

Let's first consider two test bodies at coordinate rest, both at $y = z = 0$ and separated by a small coordinate distance Δx_0 . Their physical separation is, from (11.29),

$$\Delta \ell_x = \sqrt{1 - h_{11}} \Delta x_0 = (1 - h_{11}/2) \Delta x_0. \quad (11.43)$$

(see Exercise 11.4). Thus the physical separation changes in time according to the time dependence of the function h_{11} . A useful example is to take the function to be an oscillation like a sine or cosine, so the separation of the bodies oscillates about Δx_0 by a small amount $h_{11} \Delta x_0/2$, corresponding to a fractional distance change $h_{11}/2$. Exactly the same considerations for two test bodies separated in the y direction gives us a change of $-h_{11} \Delta y_0/2$ and a fractional change of $-h_{11}/2$. In Fig. 11.1 we show the effect on a circular ring of test bodies in the x, y plane for a wave moving in the z direction. Because of the motion pattern a wave of this sort is referred to as polarized in the $+$ direction and the metric function is often written $h_{11} = h_+$. From Fig. 11.1 it is already clear how one might try to detect a gravitational wave using a distance measuring device such as an interferometer.

In the above we considered a wave with only an h_{11} component and $h_{12} = 0$. Next we will consider a wave with nonzero h_{12} and $h_{11} = 0$. There is a very easy way to do this and also display the nature of polarization for the waves. The coordinates in the x, y plane may be rotated by 45 degrees to a tilde system using the transformation

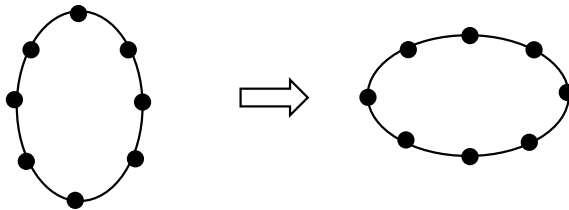


Fig. 11.1 Qualitative nature of motion produced by an oscillatory plane gravitational wave on a circle of test bodies for the $+$ polarization. The pictures are a half cycle apart. For the $\times$ polarization the pictures are rotated by 45°

$$x = \frac{1}{\sqrt{2}}(\tilde{x} + \tilde{y}), \quad y = \frac{1}{\sqrt{2}}(\tilde{x} - \tilde{y}). \quad (11.44)$$

For the line element (11.29) with $h_{11} = 0$ this gives for the tilde system

$$\begin{aligned} ds^2 &= c^2 dt^2 - dx^2 - dy^2 + 2h_{12} dx dy - dz^2 \\ &= c^2 dt^2 - (1 - h_{12}) d\tilde{x}^2 - (1 + h_{12}) d\tilde{y}^2 - dz^2. \end{aligned} \quad (11.45)$$

Since this is exactly the same line element we have just analyzed we need do no more. The motion of test bodies is the same as we obtained above but with everything rotated by 45 degrees. For this reason we refer to the waves with only nonzero h_{11} as $+$ polarized and those with only nonzero h_{12} as $\times$ polarized, and sometimes write $h_{11} = h_+$ and $h_{12} = h_\times$.

Notice an analogy between gravitational waves and electromagnetic waves. We see that the two polarizations of gravitational waves are related by a rotation of 45 degrees, whereas the 2 polarizations of electromagnetic waves are related by a rotation of 90 degrees. In the quantum field theory of these fields this is associated with the spin of a photon being 1 and the spin of a graviton being 2 (Bjorken 1965).

For people who are not interested in relativity theory, but are interested in the detection of gravitational waves, it is useful to express the dynamics of bodies in gravitational waves in terms of equivalent Newtonian tidal forces. From the expression (11.43) for the physical distance from a test body at the origin to a nearby freely falling test body we can calculate the relative velocity and acceleration to be

$$\begin{aligned} v_x &= \frac{d}{dt} \Delta \ell_x = -\frac{1}{2} \frac{d}{dt} (h_+ \Delta x_0) = -\frac{1}{2} \frac{d}{dt} (h_+ \ell_x), \\ a_x &= \frac{dv_x}{dt} = -\frac{1}{2} \frac{d^2}{dt^2} (h_+ \ell_x). \end{aligned} \quad (11.46)$$

According to Newton's second law this is the same relative acceleration that bodies would experience under a Newtonian tidal force,

$$\Delta F_x / m = -\frac{1}{2} \frac{d^2}{dt^2} (h_+ \ell_x) = -\frac{1}{2} \ddot{h}_+ \ell_x, \quad (11.47)$$

where the dot indicates a derivative with respect to time. This tells us for example that the tidal force exerted by a monochromatic gravitational wave is proportional to the square of the frequency.

$$\Delta F_x / m = \frac{1}{2} h_+ \ell_x \omega^2. \quad (11.48)$$

If one wants to design a mechanical wave detector consisting of springs and masses or solid bars this effective force can be quite useful, and it is not necessary to understand general relativity. See Exercises 11.3 and 11.4.

11.5 Gravitational Wave Sources

After solving the equations for plane gravitational waves in vacuum and seeing how they affect test bodies we now turn to understanding some possible sources of such waves. For this we use the linearized equations, which we repeat from (11.18),

$$\bar{h}_{\mu\nu}{}^{,\lambda}{}_{,\lambda} = \square^2 \bar{h}_{\mu\nu} = -\left(\frac{16\pi G}{c^2}\right) T_{\mu\nu}, \quad \bar{h}^{\mu\nu}{}_{,\nu} = 0. \quad (11.49)$$

The layout of the source region and the distant detection region that we assume is shown in Fig. 11.2.

This field equation (11.49) occurs often in physics, in particular in electrodynamics; we discuss it in Appendix 1 for the reader who is not familiar with it or desires a brief review. The retarded solution is, from (11.83) in Appendix 1,

$$\begin{aligned} \bar{h}(\vec{x}, t)_{\mu\nu} &= -\left(\frac{4G}{c^2}\right) \int \frac{1}{r} T(\vec{x}', t_{\text{ret}})_{\mu\nu} d^3x', \\ r &= |\vec{x} - \vec{x}'|, \quad t_{\text{ret}} = t - r/c. \end{aligned} \quad (11.50)$$

Here t_{ret} is referred to as the retarded time for obvious reasons. For the small source approximation, in which the source size and characteristic frequency obey $L\omega_{ch} \ll c$, the radiation from all parts of the source is in phase, and the solution far from the source reduces to an integral over the source at a single retarded time,

$$\bar{h}(\vec{x}, t)_{\mu\nu} = -\left(\frac{4G}{c^2}\right) \frac{1}{r} \int T(\vec{x}', t_{\text{ret}})_{\mu\nu} d^3x', \quad r = |\vec{x}|, \quad t_{\text{ret}} = t - r/c. \quad (11.51)$$



Fig. 11.2 The small source on the right emits gravitational waves that are to be detected at a large distance r on the left side, where they are approximately plane waves moving in the local z direction

See (11.99) in Appendix 3, for the electromagnetic analog of this.

It is interesting to observe that the solution (11.51) obtained to study gravitational waves also contains the solution for the static field (11.21) of a mass distribution, which we discussed in connection with the classical limit. To see this we set $\mu = \nu = 0$ and see immediately that $\bar{h}_{00}$ is consistent with (11.20) and (11.21) since the integral of the density is the mass M . In abbreviated notation we thus have,

$$\bar{h}_{00} = -\frac{4G}{c^2 r} \int T_{00} d^3 x' = -\frac{4GM}{c^2 r}. \quad (11.52)$$

It is always useful to have such a “sanity check.”

Equation (11.51) is a complete (but approximate) solution to the problem we posed, the metric field from a small and distant source with a known energy momentum distribution. However it has two aspects that require further attention. First, it is not in the most convenient form for typical sources, and secondly it is clearly not in the traceless transverse gauge form that is convenient for studying motion in a detector system. Dealing with these requires some straight-forward but somewhat lengthy algebra.

To begin we will show that if either of the subscripts in (11.51) is zero the integral is constant in time and thus not relevant for wave analysis. Our tool for showing this is the zero divergence of the energy-momentum tensor, that is the conservation of energy-momentum. We set $\mu = 0$ and differentiate the integral in (11.51) with respect to the retarded time to obtain

$$\frac{\partial}{\partial t} \int T^{0\nu} d^3 x' = \int T^{0\nu}{}_{,0} d^3 x'. \quad (11.53)$$

Consider first $\nu = 0$ so the zero divergence of the energy-momentum tensor implies

$$T^{0\beta}{}_{,\beta} = T^{00}{}_{,0} + T^{0i}{}_{,i} = 0, \quad T^{00}{}_{,0} = -T^{0i}{}_{,i}, \quad (11.54)$$

and, with the help of Gauss's Theorem, we may evaluate (11.53) for as a surface integral,

$$\frac{\partial}{\partial t} \int T^{00} d^3 x' = - \int T^{0k}{}_{,k} d^3 x' = - \int_S T^{0k} dS_k. \quad (11.55)$$

But the source is of limited extent by assumption, so we can choose the surface S to be outside the source region, and the surface integral is thus zero and does not correspond to gravitational waves. The same manipulations show that for $\nu = j$ the integral (11.53) is also zero. Only the space components of the energy-momentum tensor produce waves. This is consistent with our previous result (11.52).

Having disposed of the $\mu = 0$ parts of the solution (11.51) we are left with, in obvious abbreviated notation,

$$\bar{h}_{ij} = -\left(\frac{4G}{c^2}\right)\frac{1}{r}\int T_{ij}d^3x'. \quad (11.56)$$

There is a wonderful theorem that will let us calculate the integral (11.56) in a simple and physically meaningful way in terms of the quadrupole nature of the source. The theorem is

$$\int T^{k\ell}d^3x = \frac{1}{2c^2}\frac{\partial^2}{\partial t^2}\int T^{00}x^kx^\ell d^3x. \quad (11.57)$$

It is not necessary to include a prime for the space variables in the integral. As in the previous manipulations the theorem is based on the symmetry and zero divergence of the energy-momentum tensor,

$$T^{\mu\nu}{}_{,\nu} = 0, \quad \text{so} \quad T^{00}{}_{,0} = -T^{0\ell}{}_{,\ell}, \quad T^{k0}{}_{,0} = -T^{k\ell}{}_{,\ell}. \quad (11.58)$$

With the use of (11.58) and integration by parts we may evaluate the first time derivative of the integral that appears on the right side in the theorem as

$$\begin{aligned} \frac{\partial}{\partial t}\int T^{00}x^kx^\ell d^3x &= \int T^{00}{}_{,0}x^kx^\ell d^3x = -\int T^{0j}{}_{,j}x^kx^\ell d^3x \\ &= \int T^{0j}(x^kx^\ell)_{,j}d^3x = \int T^{0j}(x^k\delta_j^\ell + x^\ell\delta_j^k)d^3x. \\ &= \int (T^{0\ell}x^k + T^{0k}x^\ell)d^3x. \end{aligned} \quad (11.59)$$

In the same way we may evaluate the time derivative of the last integral above

$$\begin{aligned} \frac{\partial}{\partial t}\int (T^{0k}x^\ell + T^{0\ell}x^k)d^3x &= \int (T^{0k}{}_{,0}x^\ell + T^{0\ell}{}_{,0}x^k)d^3x \\ &= -\int (T^{kj}{}_{,j}x^\ell + T^{\ell j}{}_{,j}x^k)d^3x = \int (T^{kj}x^\ell{}_{,j} + T^{\ell j}x^k{}_{,j})d^3x \\ &= \int (T^{kj}\delta_j^\ell + T^{\ell j}\delta_j^k)d^3x = 2\int T^{k\ell}d^3x. \end{aligned} \quad (11.60)$$

It then follows from (11.59) and (11.60) that the theorem (11.57) is proved. We substitute (11.57) into (11.56) to get a simple formula for the field

$$\bar{h}_{ij} = -\left(\frac{2G}{c^4r}\right)\frac{\partial^2}{\partial t^2}\int T^{00}x'^ix'^jd^3x'. \quad (11.61)$$

The 0,0 component of the energy-momentum tensor is simply the energy density ρ , so the integral in (11.61) has a clear physical meaning and is often easy to calculate. It is generally called the quadrupole integral.

Our final manipulation is to consider the metric field (11.61) at large distances from the source, where it is asymptotically a plane wave, and to transform it to the traceless transverse gauge we used previously to calculate the motion of particles. This has already been done in general in Sect. 11.3 on plane waves. There we found that the components of the field with an index equal to 0 or 3 could be transformed away; we have also just shown that these components are constant in time and are not relevant to a wave analysis, so no more need be said about them. It only remains to transform the 1,2 block into traceless form which we already did before in (11.40). Using that equation we may write the metric field in the new system,

$$\begin{aligned} h_{11} &= -\left(\frac{G}{c^4 r}\right) \frac{\partial^2}{\partial t^2} \int T^{00}(x'^2 - y'^2) d^3 x' = -h_{22}, \\ h_{12} &= -\left(\frac{2G}{c^4 r}\right) \frac{\partial^2}{\partial t^2} \int T^{00}(x', y') d^3 x', \end{aligned} \quad (11.62)$$

in which we no longer need to label the metric field with a prime. Since it is traceless we also do not need to include the “hat” notation.

Equation (11.62) is also called the quadrupole Formula. It is in convenient form for calculation; if we know the mass density $\rho = T^{00}$ as a function of retarded time and position it gives the distant metric field in traceless transverse form. It may be applied to many real-world sources that are small and slowly moving on the astronomical scale, as we will discuss in the following examples.

Example 11.1 Some examples of the use of the quadrupole formula are in order. First consider a linear oscillator, that is a system in which all the mass is concentrated in a point oscillating along the x' axis, which is perpendicular to the z' axis. The density function is then a Dirac delta function,

$$T^{00} = \rho = M \delta(x' - R \cos \omega t) \delta(y') \delta(z'). \quad (11.63)$$

The metric field is then purely + polarized, and easily calculated from (11.62) to be

$$h_{11} = -\left(\frac{GM}{c^4 r}\right) \frac{\partial^2}{\partial t^2} (R \cos \omega t)^2 = \left(\frac{2GM}{c^2 r}\right) \left(\frac{R^2 \omega^2}{c^2}\right) \cos 2\omega t. \quad (11.64)$$

Notice that the quantities in the last two parentheses for h_{11} are dimensionless, and the second parenthesis is the square of a characteristic velocity over c . This

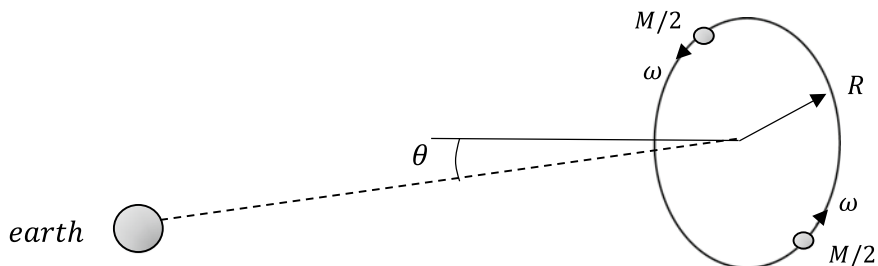


Fig. 11.3 Two bodies orbit in a plane perpendicular to the line to earth. The wave metric they produce at a distance r is given in (11.66)

is a typical form for gravitational waves and can be useful in making rough estimates. See Exercise 11.6.

Example 11.2 A more realistic example is a pair of equal mass points in circular orbit about a common center, which does occur in nature. We assume the orbit plane is fortuitously perpendicular to the earth direction in Fig. 11.3 so the geometry is simple, with $\theta = 0$. From the figure the density function is

$$\begin{aligned}
 T^{00} = \rho &= \frac{1}{2} M \delta(x' - R \cos \omega t) \delta(y' - R \sin \omega t) \delta(z') \\
 &+ \frac{1}{2} M \delta(x' + R \cos \omega t) \delta(y' + R \sin \omega t) \delta(z') \quad (11.65)
 \end{aligned}$$

From the expressions in (11.62) the metric field is then

$$\begin{aligned}
 h_{11} &= -\left(\frac{GM}{c^4 r}\right) \frac{\partial^2}{\partial t^2} [(R \cos \omega t)^2 - (R \sin \omega t)^2] \\
 &= \left(\frac{4GM}{c^2 r}\right) \left(\frac{R^2 \omega^2}{c^2}\right) \cos 2\omega t, \\
 h_{12} &= -\left(\frac{2GM}{c^4 r}\right) \frac{\partial^2}{\partial t^2} [R^2 \cos \omega t \sin \omega t] \\
 &= \left(\frac{4GM}{c^2 r}\right) \left(\frac{R^2 \omega^2}{c^2}\right) \sin 2\omega t. \quad (11.66)
 \end{aligned}$$

Thus the metric field has both $+$ and $\times$ polarizations; it is the gravitational analog of circularly polarized light.

In Exercise 11.5 you are asked to work out the wave metric for a general angle $\theta > 0$, rather than $\theta = 0$. The result is

$$\begin{aligned}
h_{11} &= \left(\frac{4GM}{c^2 r} \right) \left(\frac{R^2 \omega^2}{c^2} \right) \cos 2\omega t \left(\frac{1 + \cos^2 \theta}{2} \right), \\
h_{12} &= \left(\frac{4GM}{c^2 r} \right) \left(\frac{R^2 \omega^2}{c^2} \right) \sin 2\omega t (\cos \theta).
\end{aligned} \tag{11.67}$$

Example 11.3 (lengthy) In the above examples we ignored the important fact that the orbital system must lose energy as it emits gravitational waves. In the process of losing energy the frequency of the orbital system and the emitted waves increases; this frequency increase is crucial in the detection process. Because of the frequency increase over time the signal is generally referred to as a chirp.

In this rather long example we will only sketch the calculation of the chirp signal for the same orbital system as in the previous example shown in Fig. 11.3. We do this because the algebra involved is somewhat tedious and not very informative, and also because the calculation involves the energy content of the waves, which we have not discussed. Our goal is only to give a qualitative understanding of the chirp waveform. For the reader interested in more detail we have included Exercises 11.9–11.15. See also Schutz (1986), Holz (2019) and Kenyon (1990).

The quadrupole formula (11.62) is valid for low velocities and weak fields, and gives (11.67) for the case of a constant frequency source. Note that it has a simple qualitative form in terms of a characteristic velocity, which we mentioned in Example 11.1 and also in Exercise 11.6,

$$h \sim \left(\frac{GM}{c^2 r} \right) \left(\frac{v_{\text{char}}^2}{c^2} \right) \cos 2\omega t. \tag{11.68}$$

There is an alternative form and notation for the wave in (11.67) that is convenient and useful (Holz 2019). We continue to assume classical gravitational mechanics for the orbital system, and thus have Kepler's law giving the orbital radius R in terms of the frequency ω , and we also introduce a chirp mass M_{ch} ,

$$R^3 = \frac{GM}{8\omega^2}, \quad M_{\text{ch}} \equiv \frac{(m_1 m_s)^{3/5}}{(m_1 + m_s)^{1/5}} = \frac{M}{2^{6/5}}, \tag{11.69}$$

These two relations allow us to write the signal (11.67) in the alternative form

$$\begin{aligned}
 h_{11} &= \frac{4c}{r} T_{\text{ch}}^{5/3} \omega^{2/3} \cos 2\omega t \left(\frac{1 + \cos^2 \theta}{2} \right) \\
 h_{12} &= \frac{4c}{r} T_{\text{ch}}^{5/3} \omega^{2/3} \sin 2\omega t (\cos \theta), \quad T_{\text{ch}} = \left(\frac{GM_{\text{ch}}}{c^3} \right). \quad (11.70)
 \end{aligned}$$

The quantity T_{ch} in parenthesis is referred to as the chirp time. It is the time it takes a light signal to cross the geometric chirp mass in (11.69). See Exercise 11.10.

Our basic goal is to include the dissipation of energy in the orbital system due to radiation since that is the kind of event that has been detected to date. The orbital system radiates waves, loses energy to the waves, and spirals in. The frequency ω must therefore increase during this inspiral and thus the amplitude h must also increase according to (11.70), until the orbital system coalesces. It is rather obvious that the calculation of the system and the waves during coalescence requires numerical methods and is beyond our present scope.

One can get the energy density in gravitational waves in many ways (Kenyon 1991). One simple heuristic way is by analogy with electromagnetic waves, as in Exercises 11.11 and 11.12. The result for the energy density in a wave is

$$\rho_E = \frac{c^2}{16\pi G} \langle (\dot{h}_{11})^2 + (\dot{h}_{12})^2 \rangle \propto \omega^2. \quad (11.71)$$

Here the angle brackets mean average over a wavelength or so. The mechanical energy of the orbiting system is easy to get in terms of its frequency ω and is

$$E = -\frac{(GM)^{5/3}}{8G} \omega^{2/3}. \quad (11.72)$$

If we balance the energy lost by the orbital system in a short time with the energy given to the wave we obtain a simple equation for the frequency change

$$\frac{d\omega}{dt} = \frac{96}{5} T_{\text{ch}}^{5/3} \omega^{11/3}, \quad T_{\text{ch}} = \frac{GM_{\text{ch}}}{c^3}. \quad (11.73)$$

(see Exercise 11.13). Thus the frequency of the orbital system increases rather rapidly with time as does the wave amplitude and frequency of the wave according to (11.70).

We can solve for the time dependence of the frequency from (11.73). The solution is elementary and may be written in terms of an initial frequency ω_{in} and the elapsed time Δt after some arbitrary initial time as

$$\omega = \omega_{\text{in}} \left(1 - \frac{\Delta t}{T_{\text{co}}}\right)^{-3/8}, \quad \omega_{\text{in}} = \text{initial frequency at } \Delta t = 0,$$

$$\frac{1}{T_{\text{co}}} = \frac{256}{5} \omega_{\text{in}}^{8/3} T_{\text{ch}}^{5/3}, \quad \text{coalescence time.} \quad (11.74)$$

The first thing that (11.74) tells us is that the point masses coalesce at time $\Delta t = T_{\text{co}}$ when the frequency becomes infinite. This is of course not realistic since the bodies in the orbital system have finite size and coalesce sooner! See Exercise 11.15.

The second thing that (11.74) tells us is that the amplitude of the wave, which is proportional to the $2/3$ power of the frequency according to (11.70), increases in time according to

$$\omega^{2/3} = \omega_{\text{in}}^{2/3} \left(1 - \frac{\Delta t}{T_{\text{co}}}\right)^{-1/4}. \quad (11.75)$$

We can also calculate the phase of the waves from (11.74). We need only replace ωt in (11.70) by the integral of ωdt . The integral is elementary and the resultant phase is

$$\Phi(t) = \int_0^{\Delta t} \omega dt = \frac{8}{5} \omega_{\text{in}} T_{\text{co}} \left[1 - \left(1 - \frac{\Delta t}{T_{\text{co}}}\right)^{5/8} \right]. \quad (11.76)$$

From (11.75) and (11.76) it is clear that the wave chirp signal can tell us the chirp time and chirp mass directly—and also with some redundancy since both the amplitude and the frequency of the wave are measurable! That is, the chirp signal can identify the source as being an inspiraling orbital system.

For finite size bodies such as black holes and neutron stars there is an upper frequency limit as discussed in Exercise 11.15: it is not infinite. However for an inspiraling system near coalescence our entire treatment is not accurate since the system will become relativistic and the dynamics will be more complex. The coalescence process itself is also clearly not describable in classical terms.

Let us summarize this long example. The waveform for the inspiraling system is a chirp with the form

$$h_{11} = \frac{4c}{r} T_{\text{ch}}^{5/3} \frac{\omega_{\text{in}}^{2/3}}{\left(1 - \frac{\Delta t}{T_{\text{co}}}\right)^{1/4}} \cos 2\Phi(t) \left(\frac{1 + \cos^2 \theta}{2} \right),$$

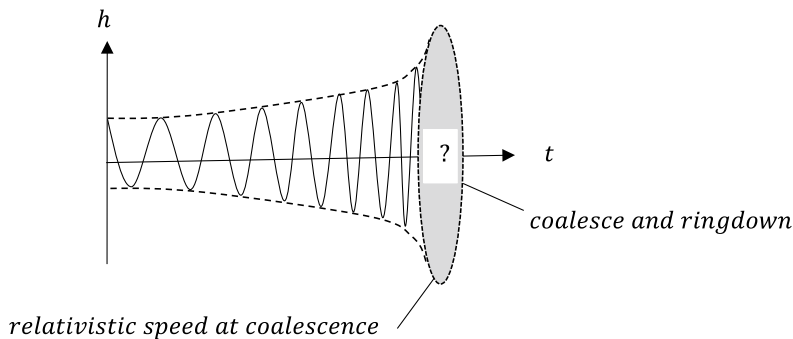


Fig. 11.4 A qualitative sketch showing the chirp signal according to the quadrupole formula and a classical analysis. The coalescence and ringdown regions are beyond the scope of this analysis

$$h_{12} = \frac{4c}{r} T_{\text{ch}}^{5/3} \frac{\omega_{\text{in}}^{2/3}}{\left(1 - \frac{\Delta t}{T_{\text{co}}}\right)^{1/4}} \sin 2\Phi(t) (\cos \theta), \quad \Phi(t) \text{ in (11.76)} \quad (11.77)$$

The shape of the wave is thus qualitatively as shown in Fig. 11.4. It may be compared to the actual waveforms actually detected and discussed in the next Sect. 11.6.

In the next section we will discuss the observations of waves from black holes and neutron stars that are much like those in the Example 11.3.

11.6 Detection of Gravitational Waves

The topic of this book is general relativity theory and the mathematics on which it is based, but we must discuss, at least briefly, the observations of gravitational waves that have brought the theory to the forefront in astronomy and astrophysics.

After the inception of general relativity in 1915 it was clear to most theorists that gravitational waves must exist, although there was a period of confusion and some skepticism, notably by Einstein himself. Only after some decades was there any solid observational evidence. The first well-known and generally accepted evidence was indirect, and concerned the orbit of the binary pulsar system PSR B1913 + 16, discovered in 1974 and known as the Hulse-Taylor system (Hulse 1975). It is a pulsar and neutron star in close orbit. The pulses from the system may be very precise timed and this allows the parameters of the orbit, such as its period, to be measured accurately. Over some decades the period has decreased, due to the energy lost to gravitational waves; the rate of decrease of the period is directly calculable

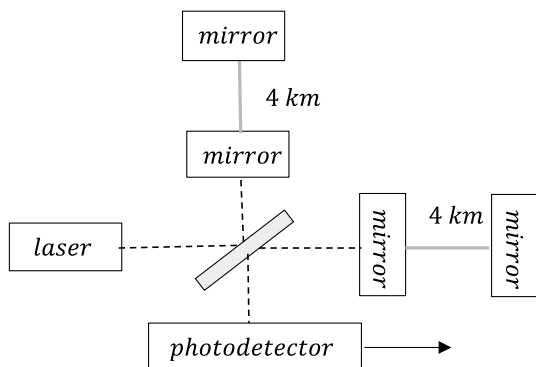


Fig. 11.5 Very simplified diagram of the LIGO Michelson interferometer layout

as we showed in Example 11.3, and the calculated value agrees quite well with the observations (Weisberg 2005).

More recently, in 2003, another system with two pulsars in orbit, PSR J0737-3039, was discovered, and it also exhibits a period decrease in good agreement with relativity theory. These systems provide some of the best observational tests of relativistic orbit predictions. They also provide excellent indirect evidence for gravitational waves, but it is clearly desirable to have more direct evidence.

It took a full century after general relativity was developed to directly detect gravitational waves. In 2015 the laser interferometric gravitational wave observatory, LIGO, made the first such detection (Abbott 2016). LIGO consists of two large interferometers placed several thousand km apart. Each is an L-shaped Michelson interferometer with 2 arms, each about 4 km in length. The mirrors at the end of each arm are suspended in such a way as to be free to move in the horizontal plane when acted upon by a gravitational wave. See Fig. 11.5. The sensitivity to motion is rather astounding, roughly a part in 10^{21} , so motion of the mirrors of much less than a proton radius can be detected, that is of order 10^{-17} m. Almost needless to say a major part of the task in developing such a detection system involves reducing the noise due to seismic and other sources to extremely low levels. To reach the needed sensitivity required decades of development.

The operation of LIGO is conceptually very simple. Take for example a wave with + polarization, such as we analyzed in Sect. 11.5, that passes in the z direction with the interferometer in the x, y plane; then the distance between the mirrors in the x direction first increases while the distance between the mirrors in the y directions decreases; then of course the motion cycles as shown in Fig. 11.1. For other orientations there are simple geometric factors to consider, as we discussed in Example 11.3. One can thereby obtain information about the direction and polarization of the source.

It is important to remember that the mirrors of LIGO are effectively free to move in response to the gravitational wave. The rest of the structure is effectively rigid due to interatomic forces.

The first signal seen by LIGO, in September 2015, was a pulse that began as a sine wave, then increased in frequency to become a chirp and ended with further decreasing oscillations. It fits very well the scenario we discussed in Sect. 11.5 and Example 11.3, as illustrated in Fig. 11.4; the signal is interpreted as coming from a pair of black holes in close orbit, losing energy by gravitational radiation, decreasing their orbital period as they move closer, and merging to form a larger black hole, which undergoes “ringdown” oscillates. Its name is GW 150914 and a sketch of its waveform is shown in Fig. 11.6a. A matching theoretical template obtained using numerical methods is shown in Fig. 11.6b. The early part of the waveform is indeed consistent with Fig. 11.4. Some parameters of the event are shown in Table 11.1 (LIGO).

It is notable that the masses of the black holes in GW150914 were rather larger than expected, about 30 solar masses, and also notable that the velocity near merger

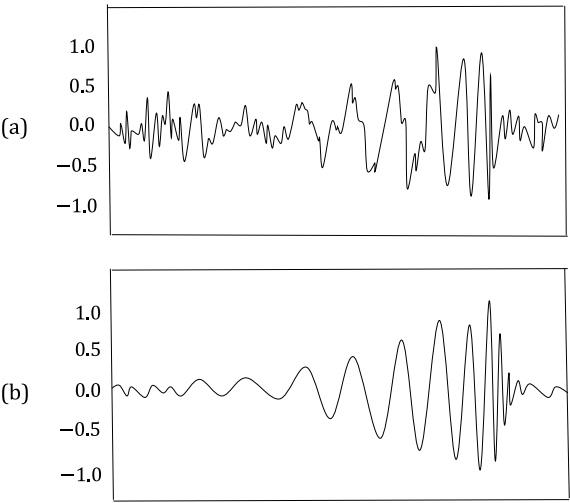


Fig. 11.6 Waveform sketch of GW150914 is shown in (a). Theoretical matching template sketch is shown in (b); the merger and ringdown are included. The vertical scale is in 10^{-21} units and the duration is about 0.5 s

Table 11.1 Parameters associated with GW150914

Distance	0.75–1.9 Gly	Peak GW strain	10^{-21}
Redshift	0.054–0.136	Radiated GW energy	$2.5\text{--}3.5\,m_{\odot}$
Signal to noise	24	Peak speed of BHs	$0.6\,c$
Total mass ($m_{\odot}$)	60–70	Duration of event	$\sim 1\,\text{s}$
Primary BH	32–41	Frequency	Ballpark of 50 Hz
Secondary BH	25–33		
Remnant BH	58–67		

was about half the velocity of light. It is fair to say that this event was the first direct evidence of truly strong gravitational effects, in which the geometry at the event differed greatly from flat during the merger.

Later parts of the signal due to the merger and the following ringdown of the remnant black hole require sophisticated numerical methods which we will not discuss. Suffice it to say that the signal in its entirety can be reasonably well calculated, and is in good agreement with the observation.

After GW150914 there have been many more black hole merger events detected by LIGO. Up-to-date information and data links can be found at the LIGO website (LIGO).

In 2017 another important type of event, GW170817, was seen by LIGO, the inspiral and merger of two neutron stars into a final neutron star (Abbott 2017). Unlike the black hole events the neutron star event produced electromagnetic radiation across the entire spectrum, from radio waves to gamma rays, and left little doubt that LIGO was truly detecting astronomical gravitational wave sources. It also verified that gravitational waves move at the same speed as light to excellent accuracy.

It is fundamentally important to realize that the variety of black hole and neutron star events as seen by LIGO constitutes an entirely new window on the universe and not just a test of the predictions of relativity. The window is likely to greatly enhance our understanding of astrophysics and the universe. For example the analysis of such events can provide an independent measurement of the Hubble constant, as we will discuss further in Part IV (LIGO 2017).

Our discussion has focused on the LIGO detector system based in the US. There is also a collaborative system named Virgo based in Italy; several more earth-based systems are expected to be operating in the near future. There are also plans for a space system, the Laser Interferometric Space Antenna or LISA, which would be millions of km in size and able to detect much lower frequencies than earth-based systems. See reference (LIGO).

Gravitational waves could in principle be produced and detected in a laboratory environment. However it is clear that this would be exceedingly difficult and thus the astrophysical sources will likely be the only sources of information for the foreseeable future; see Exercise 11.16.

Appendix 1: Solutions for Retarded Potentials

Equations involving the d'Alembertian operator and a source, such as (11.18), are ubiquitous in physics. As such, anyone who has studied electromagnetism is familiar with them (Jackson 1999). This Appendix is essentially a short reminder of the solutions and their meaning. We consider an equation that relates a field ψ via the d'Alembertian operator to some source f according to

$$\square^2 \psi(\vec{x}, t) = 4\pi f(\vec{x}, t), \quad \square^2 \equiv \eta^{\alpha\beta} \frac{\partial}{\partial x^\alpha} \frac{\partial}{\partial x^\beta} = \frac{\partial^2}{\partial t^2} - \nabla^2. \quad (11.78)$$

We will not give a rigorous derivation of the relevant solutions but instead a convincing heuristic discussion.

The time independent case is very familiar; it is the same as Coulomb's law of electrostatics; for a unit point source at the origin, $\vec{x} = 0$ and $r = 0$, the solution is

$$\psi(\vec{x}) = -\frac{1}{r}, \quad r = |\vec{x}|. \quad (11.79)$$

For a localized distribution of the source f we may superpose a continuum of such point solutions and get a more general solution

$$\psi(\vec{x}) = - \int \frac{1}{r} f(\vec{x}') d^3 x', \quad r = |\vec{x} - \vec{x}'|. \quad (11.80)$$

Such superposition is a key element in linear theories that allows relative ease of solution.

For the time dependent case we proceed in a similar way. We first look for a solution for a point source that only exists for an instant, that is the source is a delta function in space and time

$$f(\vec{x}, t) = \delta(t - t') \delta^3(\vec{x} - \vec{x}'). \quad (11.81)$$

The solution of (11.78) for such a point source is called the Green's function, and can be written

$$G(\vec{x}, t : \vec{x}', t') = \frac{1}{r} \delta(t - (t' + r/c)), \quad r = |\vec{x} - \vec{x}'|. \quad (11.82)$$

Here the time t is the time t' at the source plus the travel time to the field point. Thus a virtual point source localized in spacetime produces the same field as in the static case (11.79) but at a later time due to propagation of the effect at velocity c . A superposition of such fields gives a general solution formed by an integral, analogous to what we did in the static case. That is

$$\begin{aligned} \psi(\vec{x}, t) &= \int \frac{1}{r} f(\vec{x}', t') \delta(t - (t' + r/c)) dt' d^3 x' \\ &= \int \frac{1}{r} f(\vec{x}', t_{\text{ret}}) d^3 x', \quad r = |\vec{x} - \vec{x}'|, \quad t_{\text{ret}} = t - r/c. \end{aligned} \quad (11.83)$$

The quantity t_{ret} is called the retarded time for obvious reasons, and the above solution is called the retarded solution.

We actually have the option of choosing the opposite sign in the last equation to give an advanced time $t_{\text{adv}} = t + r/c$ and an advanced solution. It appears however that nature has chosen the retarded time: this is called causality and is usually taken as a general principle of physics.

There are two further simplifications one may often make for many sources, including masses radiating gravitational waves and charges radiating electromagnetic waves. First, if the size of the source L is much smaller than the distance r then the factor $1/r$ may be removed from the integral. Second, if the time delay for light traveling across the source, L/c , is negligible compared to the characteristic time of change for the source, call it $1/\omega_{\text{ch}}$, then the waves from each source element are in phase and the integral may be done for a single time. Then (11.83) is

$$\psi(\vec{x}, t) = \frac{1}{r} \int f(\vec{x}', t_{\text{ret}}) d^3x', \quad r = |\vec{x}|, \quad t_{\text{ret}} = t - r/c, \quad L\omega_{\text{ch}} \ll c. \quad (11.84)$$

This may be called the small source approximation.

In this Appendix we have given the Green's function or point source solution (11.82) without a rigorous derivation. Instead we chose to show the result intuitively but convincingly. For the interested reader a derivation can be obtained in a straightforward way using integrals in the complex plane, as is done in many texts on electricity and magnetism (Jackson 1999).

Appendix 2: Electromagnetic Plane Waves

We provide here a brief sketch of the solution of Maxwell's equations for plane electromagnetic waves to demonstrate how similar the mathematics is to the gravitational wave mathematics in Sect. 11.3. The theory of electromagnetic waves can be nicely expressed in terms of the 4-vector potential A_μ . It obeys the wave equation and a Lorentz gauge condition, by choice,

$$A_\mu{}^{;\lambda}{}_{;\lambda} = \square^2 A_\mu = 0, \quad (11.85a)$$

$$A^\nu{}_{;\nu} = 0. \quad (11.85b)$$

For plane waves the solution of the wave equation may be expressed as an arbitrary smooth function $\ell(U)$; here the wave vector is denoted as k_β and the quantity $U = k_\beta x^\beta$. This is easily shown by substitution, and holds for a null wave vector,

$$A_\mu = \epsilon_\mu \ell(k_\beta x^\beta), \quad k_\beta k^\beta = 0. \quad (11.86)$$

It remains to determine the polarization vector ϵ_ν , which must be consistent with the Lorentz condition (11.82). With the solution in (11.86) this condition is

$$\epsilon^\beta k_\beta = 0. \quad (11.87)$$

If we align the z axis along the space part of the null vector we may take $U = ct - z$ or some constant multiple as in Sect. 11.3. We may choose the polarization to be along either the x or y direction and thus obtain the solution in terms of either components or unit vectors as

$$A_\nu = (0, A_1(U), A_2(U), 0) = A_1(U)\hat{e}_1 + A_2(U)\hat{e}_2. \quad (11.88)$$

The picture we thereby obtain is that the polarization vector has no time component and points along either the x or y direction, perpendicular to the propagation direction z of the wave. The wave is therefore called transverse.

A transformation to what we will call a tilde gauge is defined in terms of a scalar function φ as

$$\tilde{A}_\nu = A_\nu + \varphi_{,\nu}. \quad (11.89)$$

This does not change the antisymmetric Maxwell electromagnetic field tensor, which is related to the vector potential by

$$F_{\mu\nu} = A_{\mu,\nu} - A_{\nu,\mu}. \quad (11.90)$$

Moreover, such a gauge change can take us between various choices of the polarization.

By using a gauge transformation we can put any solution of the equations (11.85) into the transverse form (11.88). In terms of $U = k_\beta x^\beta$ the solution and Lorentz gauge condition are

$$A^\mu = A^\mu(U), \quad A^\nu_{,\nu} = \frac{\partial A^\nu}{\partial x^\nu} = \frac{\partial U}{\partial x^\nu} \frac{\partial A^\nu}{\partial U} = A^\nu_{,U} k_\nu = 0. \quad (11.91)$$

Here the notation $,U$ denotes differentiation with respect to the argument U . For the gauge function φ we choose some function of U to be determined; because of (11.89) the function φ must be a solution of the wave equation,

$$\square^2 \varphi = \varphi_{,\lambda}{}^{,\lambda} = 0. \quad (11.92)$$

In the tilde system the vector potential and its divergence are then

$$\tilde{A}_v = A_v + \varphi_{,v}, \quad \tilde{A}^{\cdot v}_{\cdot v} = A^v_{\cdot v} + \varphi^{\cdot v}_{\cdot v} = A^v_{\cdot v} = 0. \quad (11.93)$$

Thus if we begin in a Lorentz gauge and make a transformation with a solution of the wave equation we remain in the Lorentz gauge; this is convenient and elegant.

Most importantly we can choose φ so that in the tilde gauge $\tilde{A}^0 = \tilde{A}^3 = 0$. For simplicity we align the vector k_v along the time and z axis, $k_v = (1, 0, 0, -1)$, so that $U = ct - z$. Then the Lorentz gauge condition is

$$A^v_{\cdot U} k_v = A^0_{\cdot U} - A^3_{\cdot U} = 0, \quad A^0 = A^3, \quad A_0 = -A_3. \quad (11.94)$$

Here the last step follows from integration, since all components are functions of only U , and any constant would be irrelevant. Exactly the same Formula holds in the tilde gauge since the Lorentz condition holds there also. Finally, in the tilde gauge we may then force the 0 and 3 components to be zero according to (11.93) by choosing

$$\begin{aligned} \tilde{A}_0 = A_0 + \varphi_{,0} = 0, \quad \varphi_{,0} = -A_0, \quad \varphi_{,U} = -A_0 \\ \tilde{A}_3 = A_3 + \varphi_{,3} = 0, \quad \varphi_{,3} = -A_3, \quad \varphi_{,U} = A_3. \end{aligned} \quad (11.95)$$

The two expressions for the U derivative of φ are same according to (11.94). Thus we can integrate to give φ as a function of U , with an irrelevant constant. Thereby the vector potential has only 1,2 components in the tilde system. Note also that the 1,2 components of the vector potential are not changed by the gauge transformation.

From the discussion of gravitational waves in the text and the above comments on electromagnetic waves the following mathematical analogies are apparent:

$$A_\mu \leftrightarrow h_{\alpha\beta} \quad \text{vector potential, metric perturbation} \quad (11.96)$$

$$\epsilon_\mu \leftrightarrow \epsilon_{\alpha\beta} \quad \text{polarization vector, metric polarization}$$

$$A'_v = A_v + \varphi_{,v} \quad \leftrightarrow \quad h'_{\alpha\beta} = h_{\alpha\beta} + (f_{\alpha,\beta} + f_{\beta,\alpha})$$

gauge transformation, small coordinate transformation

$$F_{\mu\nu} \leftrightarrow R_{\alpha\beta\gamma\delta} \quad \text{physical electromagnetic, gravitational tidal fields}$$

These analogies are rather elegant and simple. However this does not mean that electromagnetism and gravity are in any sense the same thing with a few indices altered. Einstein spent many of his later years trying to establish a deep physical connection between gravity and electromagnetism and did not succeed.

Appendix 3: Electromagnetic Wave Sources

As in Appendix 1 we give a brief sketch of the solution of Maxwell's equations with a source to demonstrate the similarity to the gravitational wave mathematics in Sect. 11.5. In terms of the 4-vector potential A_μ the equations to be solved are

$$A_{\mu}{}^{,\lambda}{}_{,\lambda} = \square^2 A_\mu = J_\mu, \quad (11.97a)$$

$$A^v{}_{,v} = 0. \quad (11.97b)$$

Notice that these are consistent with the conservation of charge relation $J^\mu{}_{,\mu} = 0$. As we discuss in Appendix 1 the retarded solution is

$$A(\vec{x}, t)^\mu = -\left(\frac{1}{4\pi}\right) \int \frac{1}{r} J(\vec{x}', t_{ret})^\mu d^3x'. \quad (11.98)$$

Far from the source we expect such a wave to approach a plane wave, and we know from our discussion of plane waves in Appendix 2 that for such a wave there is a gauge in which the 0 and 3 components of the field vanish, so we may focus on the 1,2 components of the field. Moreover, if we assume the small source approximation we may remove the $1/r$ factor from the integral and evaluate the integral at a single retarded time to obtain

$$A(\vec{x}, t)^k = -\left(\frac{1}{4\pi}\right) \frac{1}{r} \int J(\vec{x}', t_{ret})^k d^3x',$$

$$r = |\vec{x}|, \quad t_{ret} = t - r/c, \quad L\omega_{ch} \ll c. \quad (11.99)$$

For a “sanity check” note that the zeroth component of this expression is Coulomb's law involving the total charge. Thus the problem reduces to finding integrals over the space components of the current.

The integral on the right side of the solution (11.99) may be simplified with the use of the conservation of current relation, which states that the divergence of the current is zero. That is

$$J^\mu{}_{,\mu} = 0, \quad J^0{}_{,0} = -J^i{}_{,i}. \quad (11.100)$$

From this we see that

$$\frac{\partial}{\partial t} \int J^0 x'^k d^3x' = \int J^0{}_{,0} x'^k d^3x' = - \int J^i{}_{,i} x'^k d^3x' = \int J^k d^3x'. \quad (11.101)$$

Thus the integral reduces to the time derivative of an integral involving the charge density J^0 that we will call a dipole integral.

$$A(\vec{x}, t)^k = -\left(\frac{1}{4\pi}\right)\frac{1}{r}\frac{\partial}{\partial t}\left(\int J^0 x'^k d^3x'\right). \quad (11.102)$$

This result lets us calculate some simple and interesting cases of radiation. For example consider a point charge q oscillating along the x axis with amplitude L at frequency ω . Then the charge density function is

$$J^0 = q\delta(x' - L\cos\omega t)\delta(y')\delta(z'), \quad (11.103)$$

and the field from (11.102) is

$$A(\vec{x}, t)^1 = \left(\frac{q}{4\pi}\right)\frac{L\omega}{r}\sin\omega t. \quad (11.104)$$

This is a reasonable model for a short dipole antenna. Note that the corresponding electric field is the derivative of this vector potential and thus is proportional to the square of the frequency.

Exercises

- 11.1 Write out the approximate metric in (11.22) for a source which has monopole and quadrupole moments. What of a dipole moment? What of higher moments? Where might this equation be useful?
- 11.2 Do the two functions in the traceless transverse gauge solution (11.28) need to be related to each other? Can you imagine a source in which the elements of the metric have different time dependence?
- 11.3. Design a simple gravitational wave detector using springs and masses. Design one consisting of elastic rods. Equations (11.47) and (11.48) should be a help.
- 11.4 The current official definition of physical distance is that of light travel time. Consider then the line element (11.29). Light moves on a null line, $ds = 0$, at constant physical velocity c , so in the x direction it obeys $cdt = (1 - h_{11}/2)dx$. The definition of distance thus means the relation between coordinate distance and physical distance is $d\ell = cdt = (1 - h_{11}/2)dx$. This gives justification for the relation (11.43) for test bodies in a gravitational wave. Now use this to analyze the operation of an interferometer wave detector that is not of negligible size compared to the wavelength of the gravitational wave. This analysis is relevant for very large machines such as LISA.
- 11.5 We studied in Example 11.2 an orbiting pair of equal mass bodies in a plane perpendicular to the line toward earth; see Fig. 11.3. Work out the metric field if the line to earth is at an angle θ from the perpendicular, and thereby

show that the $+$ polarized wave picks up a factor of $(1 + \cos^2\theta)/2$ and the $\times$ polarized wave picks up a factor of $\cos\theta$.

Thus verify (11.67).

- 11.6 The amplitude solution (11.64) has a general order of magnitude form involving a characteristic velocity v_{ch} and we may write it as

$$h \sim \left(\frac{GM}{c^2 r} \right) \left(\frac{v_{\text{ch}}^2}{c^2} \right) = \left(\frac{m}{r} \right) \left(\frac{v_{\text{ch}}}{c} \right)^2.$$

Can you show heuristically that this should be roughly true for a fairly general source? See also Exercise 11.16.

- 11.7 In the text we discussed as sources of gravitational waves orbiting black holes and neutron stars. Can you think of any other possibly interesting astronomical sources?
- 11.8 In Sect. 11.2 on Newtonian limits we neglected some velocity dependent effects in the geodesic motion of particles and also velocity dependent effects on the sources of gravity. Such effects are interesting, although usually very small in the real world. They are called gravitomagnetic effects or sometimes “frame dragging” effects; they have been observed in the orbits of satellites and on the precession of an orbiting gyroscope. Work out the effects for the field produced by a spinning body and on the motion of bodies; see Adler (2000).
- 11.9 Verify Kepler’s law expressed in (11.69) for the orbiting system of Example 11.3.
- 11.10 Verify that the wave metric may be written in terms of the chirp time as in (11.70). What is the approximate chirp time in seconds for a pair of orbiting solar mass black holes?
- 11.11. It is fairly straight-forward to derive the energy density in a gravitational wave using linearized general relativity, but a bit tedious and lengthy. We may take a shortcut and obtain the result heuristically using the analogy with electromagnetic waves and dimensional analysis. The energy density in an electric field is well-known by physics students to be proportional to $\vec{E}^2$ and for an electromagnetic wave it is thus proportional to $(\dot{A})^2$. Use the analogy between electromagnetism and linearized general relativity discussed in Appendix 2 and (11.96) to see that the analogous expression for the gravitational wave should have the form

$$\rho \propto \langle (\dot{h}_{11})^2 + (\dot{h}_{12})^2 \rangle.$$

The angle brackets in this expression indicate that the quantity is to be averaged over a wavelength or so. Think a bit about why the averaging should occur and see Schutz (2009).

- 11.12 Next use dimensional analysis to see that a factor of c^2/G should be included in the expression for the energy density in the above exercise. The energy

density thus becomes

$$\rho = \frac{1}{16\pi} \frac{c^2}{G} \langle (\dot{h}_{11})^2 + (\dot{h}_{12})^2 \rangle$$

where the numerical factor $1/16\pi$ must be gotten from a more detailed analysis such as Schutz (2009).

- 11.13 Verify the expression (11.72) for the total energy of the orbiting system according to classical mechanics. Then use (11.70) and (11.71) to calculate the wave energy in a thin spherical shell of thickness cdt . Balance this energy with the energy the orbiting system must lose in a time dt to verify (11.73).
- 11.14 Verify (11.74) which gives the frequency of the waves as a function of time.
- 11.15 Take the size of the orbiting bodies in Example 11.3 to be nonzero and calculate a more realistic coalescence time than given by (11.74). Also calculate the maximum frequency due to the finite size.
- 11.16 We noted in the text that the production and detection of gravitational waves in a laboratory is not likely in the foreseeable future. Use the order of magnitude relation in Exercise 11.6 to show this; you need only estimate the maximum mass and velocity one might hope to achieve in a terrestrial laboratory setting.

Part IV

Cosmology

Cosmology seeks to answer the immodest question “What is the universe?” (Freedman 2006). Cosmology entered the mainstream of physics only in the last half of the twentieth century. This is due firstly to the existence of a viable theoretical structure in general relativity. Secondly, it is due to diverse observations of distant galaxies and measurements of the cosmic microwave background radiation that have confirmed the basic theoretical ideas and provided new questions. Observational cosmology has become diverse and sophisticated. It is an active part of science and no longer largely speculative. There are many questions remaining, some very deep, and most cosmologists would agree that the field is still in its infancy.

Our convention concerning the word “universe” should be noted here. For the real-world universe, some authors capitalize U. For our theoretical or conceptual universe, some authors use a small u. To avoid confusion, we will never capitalize the “u” and hope we will be forgiven for appearing to demote the real world.

We will set up the Einstein field equations in Chap. 12 and study the way in which they relate to the constituents of the universe. In Chap. 13, we will obtain the cosmological metric based on general considerations of symmetry and show some of its consequences. In Chap. 14, we will obtain the specific dynamical equations that the cosmological metric obeys. In Chap. 15, the dynamical equations will be solved for some interesting models, in particular the standard or Λ CDM model. Chapter 16 will be devoted to obtaining some of the important properties of the current universe. Finally, in Chaps. 17–19 the earlier universe will be studied, including some of its less well-understood properties and some basic unanswered questions.

Chapter 12

The Einstein Field Equations for Cosmology



Abstract Cosmology seeks to answer the rather grandiose question “What is the universe?” After about half a century of being mainly a theoretical and mathematical subject cosmology entered the mainstream of physics only in the last half of the twentieth century thanks to a viable theoretical structure provided by general relativity and diverse observations of distant galaxies and measurements of the cosmic microwave background (CMB) radiation. In this chapter we begin our study of cosmology by applying general relativity to the entire universe, in which the source of gravity is taken to be a cosmic fluid. The gravitational field equations naturally allow for one component of the fluid to be the “dark energy” that has been found by observation to be the dominant component.

12.1 The Field Equations and Energy-Momentum Conservation

In Chap. 8 we obtained the Einstein equations for the gravitational field. We also mentioned the simplest energy-momentum tensor of interest, the so-called dust tensor that describes a fluid characterized by only a mass density and velocity. In particular the dust fluid has no pressure. In this chapter we want to include fluid pressure to describe the large-scale universe of cosmology (Adler 1975; Schutz 2009). That is we want the appropriate cosmological energy-momentum tensor. We must also include the cosmological constant that describes the so-called dark energy that appears to pervade the universe and which has a large effect on its dynamics.

Recall from Chap. 8 that the field equations relate the Einstein tensor to the energy-momentum tensor of the source by

$$G^{\mu\nu} = R^{\mu\nu} - \frac{1}{2}g^{\mu\nu}R = CT^{\mu\nu}, \quad C = -\frac{8\pi G}{c^2}. \quad (12.1a)$$

Since the Einstein tensor has a zero divergence the energy-momentum tensor must also have a zero divergence, so we include a subsidiary condition,

$$G^{\mu\nu}{}_{;v} = 0, \quad T^{\mu\nu}{}_{;v} = 0. \quad (12.1b)$$

The specific dust energy-momentum tensor that we have already discussed is built from the scalar density and the 4-vector fluid flow field, and is explicitly

$$T^{\alpha\beta} = \rho u^\alpha u^\beta, \quad u^\beta = \frac{dx^\beta}{ds}. \quad (12.2)$$

In Chap. 8 we briefly discussed the implication of the zero divergence for conservation of mass but we wish to elaborate on it here and show that it also includes conservation of momentum. To illustrate this in a simple way we first study the flat space and low velocity or classical limit.

The position $x^\mu(s)$ of a particle in the dust fluid may be taken to be a function of its proper time with the 4-velocity given in (12.2). Then the 4-velocity flow of the dust to first order in velocity, as given in Chap. 2, is

$$u^\beta = \frac{dx^\beta}{ds} = \frac{1}{c} \frac{dx^\beta}{d\tau} = \frac{1}{c} \gamma(c, \vec{v}) = \left(1, \frac{\vec{v}}{c}\right) + O\left(\frac{v^2}{c^2}\right),$$

$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}}, \quad v \ll c. \quad (12.3)$$

The 4-vector velocity we use here, as defined in (12.3), is dimensionless and the proper time interval is related to the line element by $dc\tau = ds$. Notice that the 0, 0 component of the energy momentum tensor is the mass density, or energy density divided by c^2 , and the 0, i components are the momentum densities divided by c , so the name energy-momentum tensor is appropriate.

We now write out the zero-divergence condition (12.2) in this limit. For the time component $\mu = 0$

$$T^{0v}{}_{;v} = T^{0v}{}_{,v} = T^{00}{}_{,0} + T^{0i}{}_{,i} = \frac{1}{c} \left[\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x^i} (\rho v^i) \right] = 0, \quad (12.4)$$

where as usual the Latin letter i is a space index and we use standard derivative notation. This equation says that the increase of mass density ρ in a small region is balanced by the mass flow out of the region, or the divergence of ρv^i . Mass is conserved according to (12.4) as we already mentioned in Chap. 8. Similarly for a space index $\mu = j$ we have

$$T^{jv}{}_{;v} = T^{jv}{}_{,v} = T^{j0}{}_{,0} + T^{jk}{}_{,k} = \frac{1}{c^2} \left[\frac{\partial (\rho v^j)}{\partial t} + \frac{\partial}{\partial x^k} (\rho v^j v^k) \right] = 0. \quad (12.5)$$

(Terms of third order and higher in v/c are neglected.) Use of the product rule and rearrangement gives

$$\begin{aligned} & v^j \frac{\partial \rho}{\partial t} + \rho \frac{\partial v^j}{\partial t} + v^j \frac{\partial}{\partial x^k} (\rho v^k) + (\rho v^k) \frac{\partial v^j}{\partial x^k} \\ &= v^j \left[\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x^k} (\rho v^k) \right] + \rho \left[\frac{\partial v^j}{\partial t} + v^k \frac{\partial v^j}{\partial x^k} \right] = 0. \end{aligned} \quad (12.6)$$

But the first bracket of this is zero by conservation of mass in (12.4), so the second bracket is also zero and we find

$$\frac{\partial v^j}{\partial t} + v^k \frac{\partial v^j}{\partial x^k} = \frac{\partial v^j}{\partial t} + \frac{\partial x^k}{\partial t} \frac{\partial v^j}{\partial x^k} \equiv \frac{dv^j}{dt} = 0, \quad \text{cons. of momentum.} \quad (12.7)$$

This last equation states that if the velocity field is viewed as a function of time $v^j(t, x^k(t))$ then the total time derivative, or Euler derivative, of the flow velocity as defined in (12.7) is zero. That is, an element of fluid is not accelerated and its momentum is conserved.

In summary we see that the zero-divergence condition on the dust energy-momentum tensor may be interpreted in the classical limit as expressing conservation of energy and momentum. We naturally generalize this to the Riemann space of general relativity, by saying that the zero divergence of the energy-momentum tensor (12.2b) implies conservation of energy-momentum. Any source of gravity in the Einstein equations has a zero divergence because the Einstein tensor does, and its energy and momentum are thus conserved.

This property of energy-momentum conservation must be considered an extraordinarily elegant feature of general relativity. It is consistent with a fundamental assumption of the theory that gravity couples to everything that has energy and momentum.

12.2 Field Equations and the Cosmic Fluid Source

So far we have used only the energy-momentum tensor of dust, that is a fluid characterized by only a mass density and a flow velocity, and in particular with no internal pressure. A perfect fluid is more general and more realistic, in that it also has an internal pressure, but no viscosity or other fluid properties. Many real systems are rather well described as perfect fluids, for example the diffuse “gas” of galaxies that makes up the present universe on a cosmological scale, the electromagnetic radiation that dominated the universe in its early years, and the quark-gluon plasma that dominated it in its early seconds. We will discuss the energy-momentum tensor for a perfect fluid in this section. As in the previous section we will make use of the classical limit for clarity and simplicity.

In the presence of a pressure gradient an element of fluid will feel a force and be accelerated, so its momentum will change. We therefore expect that the conservation of momentum (12.7) should be modified to have a pressure gradient on the right side. In fact the equation should become Newton's second law for a fluid with density ρ in the classical limit

$$m\vec{a} = \vec{F} \rightarrow \rho \frac{d\vec{v}}{dt} = -\nabla p. \quad (12.8)$$

The time derivative is again the Euler derivative used in (12.7). (See Exercise 12.1). This is the fundamental force equation of classical fluid flow. Our task in this section is to modify the energy-momentum tensor so that it yields the dynamical equation (12.8) replacing (12.7) for a zero-pressure fluid. This will lead to the energy-momentum tensor to be used in the following chapters.

Our approach is thus to add to the energy-momentum tensor for dust a pressure term which will give the dynamical equation (12.8) in the classical limit. The procedure is analogous to that which led to (12.7). The obvious quantities available to construct such a tensor are the density ρ and pressure p , which we assume are scalars, the tensor $u^\alpha u^\beta$, and the metric tensor $g^{\alpha\beta}$. We accordingly assume the energy-momentum of the perfect fluid is

$$T^{\alpha\beta} = \rho u^\alpha u^\beta + p(au^\alpha u^\beta + bg^{\alpha\beta}) = (\rho + ap)u^\alpha u^\beta + bpg^{\alpha\beta} \quad (12.9)$$

with a and b constants to be determined. In the flat space limit the metric tensor is $g^{\alpha\beta} = \eta^{\alpha\beta}$. As with the dust tensor we calculate the divergence of this tensor and set it equal to zero as in (12.2). For $\alpha = 0$ we get an equation quite analogous to the conservation of mass (12.4).

$$\begin{aligned} T^{0v}{}_{,v} &= T^{00}{}_{,0} + T^{0i}{}_{,i} \\ &= \frac{1}{c} \left[\frac{\partial}{\partial t}(\rho + (a+b)p) + \frac{\partial}{\partial x^i}((\rho + ap)v^i) \right] = 0. \end{aligned} \quad (12.10)$$

It is immediately clear that for this to be consistent with the conservation of mass or energy in (12.4) we must have both $(a+b)p$ and ap much less than ρ .

For $\alpha = j$ we get an equation analogous to the momentum equation (12.5),

$$T^{jv}{}_{,v} = \frac{1}{c^2} \left[\frac{\partial}{\partial t}((\rho + ap)v^j) + \frac{\partial}{\partial x^k}((\rho + ap)v^j v^k - bpc^2 \delta^{jk}) \right] = 0. \quad (12.11)$$

The same manipulations with the product rule as used with (12.5) gives

$$v^j \frac{\partial}{\partial t}(\rho + ap) + (\rho + ap) \frac{\partial v^j}{\partial t} + v^j \frac{\partial}{\partial x^k}((\rho + ap)v^k) + (\rho + ap)v^k \frac{\partial v^j}{\partial x^k}$$

$$\begin{aligned}
&= v^j \left[\frac{\partial}{\partial t} (\rho + ap) + \frac{\partial}{\partial x^k} ((\rho + ap)v^k) \right] + (\rho + ap) \left[\frac{\partial v^j}{\partial t} + v^k \frac{\partial v^j}{\partial x^k} \right] \\
&= bc^2 \frac{dp}{dx^j}.
\end{aligned} \tag{12.12}$$

Let us consider (12.10) and the last line of (12.12) for a moment. The first bracket in the last line of (12.12) is approximately equal to the bracket in (12.10), which is zero. Since we know that both a and b are small we take that first bracket to be zero to a good approximation and find the sort of equation we are seeking; that is the Euler derivative of the fluid velocity represents acceleration and is proportional to the pressure gradient,

$$(\rho + ap) \left[\frac{\partial v^j}{\partial t} + v^k \frac{\partial v^j}{\partial x^k} \right] = (\rho + ap) \left[\frac{dv^j}{dt} \right] = bc^2 \frac{dp}{dx^j}. \tag{12.13}$$

Indeed if we choose $b = -1/c^2$ this has the same form as the classical fluid flow (12.8). Moreover if we also choose $a = -b = 1/c^2$ then the mass energy relation (12.12) tells us that the conserved quantity is the mass density ρ .

Let us summarize the above results: in the flat space limit if we choose the energy-momentum tensor to be

$$T^{\alpha\beta} = \rho u^\alpha u^\beta + \frac{P}{c^2} (u^\alpha u^\beta - g^{\alpha\beta}), \tag{12.14}$$

then the zero-divergence condition in that limit leads to conservation of mass and also to Newton's force equation for the fluid flow. Having verified the correctness of (12.14) in the classical limit we generalize and adopt it to represent a perfect fluid in the general case, that is with gravity and arbitrary velocities.

The added term in (12.14), proportional to pressure, is related to an object known as a stress tensor in classical continuum physics; its divergence represents a force. Thus the source tensor in the field equations could more accurately be called the energy-momentum-stress tensor, but the name energy-momentum tensor is now standard.

The perfect fluid energy-momentum tensor is characterized by only the three properties of energy density, pressure and flow velocity. The kinetic theory of gases can tell us something about the relation of the density and pressure. In the above discussion we saw that if the pressure over c^2 is much smaller than the density there is consistency with the classical limit. We can see explicitly how this comes about for the special case of an ideal gas. Recall that according to the kinetic theory of an ideal gas the pressure is given in terms of the density ρ and the root-mean-square (rms) velocity v of the gas molecules by

$$p = \rho \frac{v^2}{3}. \tag{12.15}$$

Thus the relation between pressure and density for such a gas is

$$\frac{p}{c^2} = \frac{1}{3} \left(\frac{v}{c} \right)^2 \rho. \quad (12.16)$$

Also recall that the average kinetic energy of a molecule of mass m is related to the temperature T of the gas times Boltzmann's constant k by

$$m \frac{v^2}{2} = \frac{3}{2} kT. \quad (12.17)$$

Hence for a cold gas with low velocity molecules $p/c^2 \ll \rho$, for a hot gas with high velocity molecules $p/c^2 = \rho/3$, and also for a gas of photons $p/c^2 = \rho/3$. In the present universe the gas of galactic “molecules” is quite cold, while for the early universe of high energy particles and photons the gas was very hot (see Exercises 12.2 and 12.3).

It is now standard practice to describe the fluid of the universe in terms of a parameter $w = p/\rho c^2$, giving a phenomenological equation of state of the fluid. We will discuss this at length in a later chapter.

12.3 The Cosmological Constant as Vacuum or Dark Energy

We do not yet have the most general field equations for general relativistic gravity and cosmology. The general structure of the field equations in (12.1) sets the symmetric second rank Einstein tensor representing geometry equal to the symmetric second rank tensor representing the energy-momentum content of space. The divergence of the Einstein tensor is identically zero as we have shown; the energy-momentum tensor is thus always assumed to have zero divergence, corresponding to conservation of energy-momentum. However the metric tensor also has a zero covariant derivative as we showed in Chap. 6, so it also has a zero divergence. It is thus evident that we may consistently add another term to the geometric side of the field equations, a constant multiple of the metric tensor; such a term is symmetric, second rank, and has zero divergence, so the equations remain mathematically consistent. The generalized field equations then become

$$G_{\mu\nu} + \Lambda g_{\mu\nu} = C T_{\mu\nu} = -\frac{8\pi G}{c^2} T_{\mu\nu}. \quad (12.18)$$

The added term is called the cosmological term and Λ is called the cosmological constant; it has the dimension of an inverse distance squared.

Since the field equations without the cosmological term reduce to the classical Newtonian equations it is clear that the cosmological term cannot have a large effect on the scale of the solar system. Its effect on a cosmological scale however may be

significant, and must be determined by observation. One might guess, on the basis of dimensional analysis, that its value might be comparable to the inverse square of the size of the universe.

Some theorists, notably Einstein who invented it, have objected to the cosmological term on esthetic grounds: the field equations are simpler without it. The dominant viewpoint at present is that its nonzero observed value makes it quite important; the present standard model of cosmology includes it as a major ingredient of the universe. We will discuss this further in following chapters.

The introduction of the cosmological constant in the above was by purely formal mathematical means: it is allowed by the mathematical structure of the equations. There is however an alternative physical interpretation of the cosmological term that is interesting. If we simply move the cosmological term to the right side of the field equations,

$$G_{\mu\nu} = C \left(T_{\mu\nu} - \frac{\Lambda}{C} g_{\mu\nu} \right), \quad (12.19)$$

then we may interpret it as a contribution to the total energy-momentum tensor. In the absence of any ponderable material it may be thought of as the energy-momentum tensor of empty space, that is of the vacuum. However the cosmological term corresponds to a peculiar energy-momentum tensor. Comparison with the perfect fluid energy-momentum tensor in (12.14) shows that it may be viewed as a perfect fluid if

$$-\frac{\Lambda}{C} g^{\alpha\beta} = \rho u^\alpha u^\beta + \frac{p}{c^2} (u^\alpha u^\beta - g^{\alpha\beta}). \quad (12.20)$$

This is only consistent if we take the mass density and pressure of the vacuum to be

$$\frac{p}{c^2} = -\rho, \quad \rho = -\frac{\Lambda}{C} = \frac{c^2}{8\pi G} \Lambda, \text{ mass density.} \quad (12.21)$$

If instead we use the energy density ρc^2 these look a bit simpler

$$p = -\rho_V, \quad \rho_V = \frac{c^4}{8\pi G} \Lambda, \text{ energy density.} \quad (12.22)$$

That is, the pressure is the negative of the energy density, much in contrast to the situation for an ideal gas in which the pressure is positive and smaller than the energy density. The vacuum fluid is thus quite peculiar and is now an important part of present cosmological theory. Because it does not interact directly with light it is widely called dark energy. This name also allows a more general view of its nature; the dark energy is presently the subject of intense observational and theoretical study (Amendola 2010).

With the interpretation of the cosmological term as dark energy one might reasonably argue that it should be zero: why should the vacuum, empty space, have an energy density? This viewpoint resonates with the esthetic criterion, that the field equations be as simple as possible. However in quantum field theory the vacuum does not have zero energy, and the energy density of “empty space” is not zero; instead it is “formally infinite,” and even if allowance is made for a reasonable granularity of space on a very small scale it is absurdly large. It is so large that the vacuum energy in a volume the size of a nucleus is about equal to the total energy content of the observed universe. Stated in another way the estimated “theoretical estimate” of the cosmological constant, with spacetime granularity, is about 10^{120} times the value allowed by present astronomical observations. This absurd result is called by some theorists the problem of the cosmological constant, and by others the vacuum catastrophe (Adler 1995). For those interested in reconciling quantum theory and general relativity it is a crucial problem. For those mainly interested in observationally verifiable cosmology the problem is of less importance.

Many theorists have suggested that a cosmic field of some sort could behave like the cosmological constant but have a dynamical origin; some such fields have been termed quintessence and are being actively studied, especially regarding their observable properties (Amendola 2010).

In the following chapters we will be somewhat unconventional and variously use the names cosmological constant or vacuum energy or dark energy to refer to the same generic thing.

12.4 Summary

For our further study of cosmology the field equations will be taken to be those in (12.18) with the energy-momentum tensor being that of a perfect fluid (12.14), generally called the cosmic fluid. These were motivated using classical ideas and the density ρ was taken to be a mass density. However for relativistic cosmology it is usually more convenient to use the energy density as in (12.22), $\rho_e = \rho c^2$, and thereby give the energy-momentum tensor the units of energy density, so the fundamental gravitational equations become

$$G_{\mu\nu} + \Lambda g_{\mu\nu} = CT_{\mu\nu} = -\frac{8\pi G}{c^4}T_{\mu\nu}, \quad T^{\alpha\beta} = \rho_e u^\alpha u^\beta + p(u^\alpha u^\beta - g^{\alpha\beta}). \quad (12.23)$$

Here the cosmological constant is on the left side, and there is no dark energy on the right side.

As we mentioned in Sect. 12.2 it is now standard practice to use an effective equation of state for the cosmic fluid using the parameter $w = p/\rho_e$. The value of w is 0 for cold matter, $1/3$ for hot matter or photons, and -1 for dark energy. The question of whether the cosmological term should be best thought of as part of the

geometry on the left side of the field equations, or as dark energy on the right side of the equations, is an important question in that it could determine our mindset in developing future theories.

Exercises

- 12.1 Carefully apply Newton's second law to an element of a fluid and derive the basic fluid equation (12.8).
- 12.2 Estimate very roughly the relation of pressure to density for the present-day cosmological gas of galaxies by looking up the approximate random motion velocity of a typical galaxy and using simple kinetic theory. Is the neglect of pressure in the present-day universe justified?
- 12.3 Suppose that the dominant ingredients of the present-day universe are not really the visible galaxies and gas and dust that we observe but instead are unseen very low mass and fast-moving particles such as neutrinos. What happens to your answers to Exercise 12.2? This is called the hot dark matter universe.
- 12.4 We have called Λ the cosmological constant. Could it instead be a variable depending on time, $\Lambda(t)$? Assume that energy-momentum is conserved and show that Λ must be constant by taking the divergence of both sides of the field (12.18).
- 12.5 Suppose the contrary to Exercise 12.4, that the cosmological term does vary slowly with time. What would that specifically imply for conservation of energy in the universe and what experiment or observation could test for it?
- 12.6 Could it be that some special type of energy-momentum does not couple to gravity? Show that this would not be consistent with conservation of energy-momentum.
- 12.7 As we have indicated the cosmological term can be thought of as part of the geometry on the left side of the field equations or as some sort of dark energy stuff on the right side—that is “geometry versus stuff.” What is your preference—based purely on esthetic and philosophical grounds? There is, of course, no right or wrong answer to this question.
- 12.8 Consider a cylinder filled with a material whose total energy is proportional to its volume. Using standard energy arguments as in thermodynamic textbooks show that the equation of state must be $p = -\rho c^2$ as in (12.21); that is the pressure must be negative.

Chapter 13

Cosmological Preliminaries



Abstract Observations of the universe on the largest scale of billions of light years indicate that it is expanding, is filled with cosmic microwave background (CMB) radiation, and is approximately homogeneous. These facts motivate the choice of an appropriate form of metric called the FLRW metric (Friedmann, Lemaitre, Robertson, and Walker), which we will derive in this chapter. The FLRW metric contains a fundamental function describing the expansion of the universe, called the scale factor. The FLRW metric leads to an elegant description of some physical properties of the expanding universe, such as cosmic horizons. One particular example of an FLRW metric is that of de Sitter, which is mainly of theoretical and mathematical interest.

13.1 Basic Observations and Assumptions

We begin our study of cosmology with three basic observational facts related to the observed universe on a very large scale. By that we mean a scale of billions of light years, whereas the distance between galaxies is only some millions or tens of millions of light years.

- A. *The Universe is expanding.* Distant galaxies are observed to have spectra which are Doppler shifted to the red, indicating that they are receding from us. This was first discovered by Hubble in 1929, and has been quite well confirmed since then (Hubble 1929). The velocity of recession v of relatively nearby galaxies is observed to be approximately proportional to their distance L from us, which is known as Hubble's law.

$$v = H_0 L, \quad \text{Hubble's law,}$$
$$H_0 = 70 \pm 5 (\text{km/s})/\text{Mpc} \quad \text{Hubble's constant—our error estimate.} \quad (13.1)$$

The original rough data on which this relation is based is shown in Fig. 13.1; it has been superseded by much more accurate data, so Fig. 13.1 is only of historical

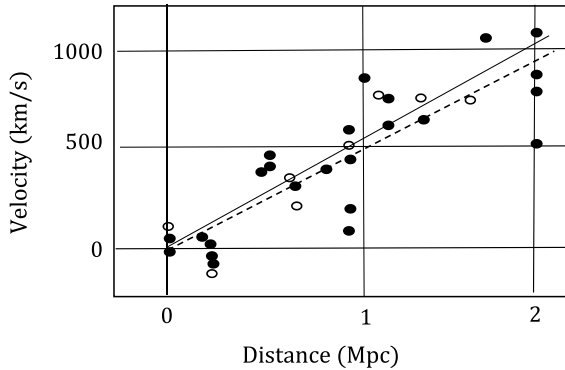


Fig. 13.1 Hubble's original data. Black dots represent individual galaxies and the solid line is a best fit to these. Open circles represent groups of galaxies and the dashed line is a best fit to these. See Fig. 13.2 for more recent data at larger distances

interest; in particular the distances are not accurate. Figure 13.2 shows more recent data based on the use of type 1a supernovae (SN 1a) as standard candles rather than entire galaxies (Kirshner 2004).

The Hubble constant is generally considered to be the present value of a slowly varying function of time, as we will study in detail. Its presently measured value using Hubble's technique is about $H_0 = 74 \text{ (km/s)/Mpc}$, with distances in megaparsecs as used by astronomers, or $1/(14 \times 10^9 \text{ year})$ in light years; we will say more about this value below and in Appendix 1 (Reiss 2019).

If we ponder Hubble's law for a moment we see that it tells us that all the galaxies and their stars would have been much closer together at about 14×10^9 year ago.

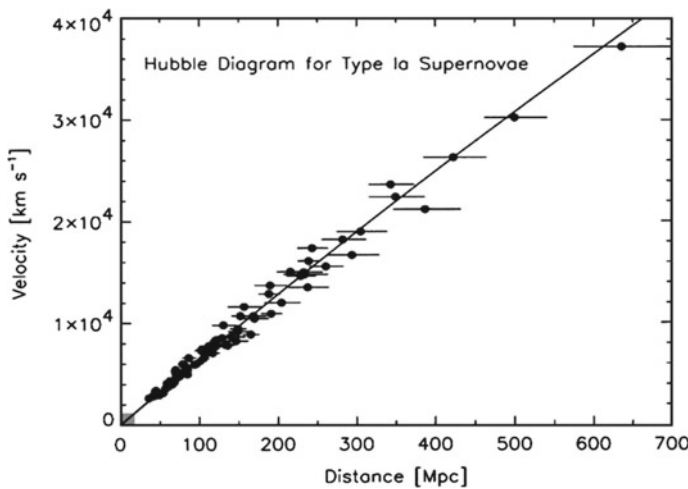


Fig. 13.2 The Hubble diagram for type 1a supernovae out to about 700 Mpc (Kirshner 2004)

Thus $1/H_0$ is a characteristic time scale for expansion of the universe and c/H_0 is a characteristic distance scale. The scenario in which all the material in the universe was close together in the past is now very well-known as the big bang, and is a fundamental part of cosmology.

There was once large uncertainty in the value of the Hubble constant. Early estimates from galaxy observations varied over the range 50 – 100 (km/s)/Mpc because of the difficulties involved in measuring cosmological distances. Because of this, astronomers have often expressed results which depend on H_0 in terms of a dimensionless number h_0 , defined by $H_0 = h_0 100$ (km/s)/Mpc. As noted above the value of the Hubble constant is now much better determined, but there remain important questions regarding its value. In particular, another method of obtaining H_0 using the CMB discussed below, gives a value that is not quite consistent with the classic Hubble method; this is discussed below and in Appendix 1 and also in Chap. 17 (Reiss 2019; Planck 2018; Chalinor 2012).

The linear Hubble law (13.1) only holds for galaxies relatively close to us on the cosmological scale. For more distant objects, galaxies and supernovae, observations show a deviation from linearity, and this gives information on the material of the universe. Specifically the data indicate that the universe is not only expanding but that the expansion is accelerating, as we will later discuss (Reiss 1998; Perlmutter 1999; Kirshner 2004). This in turn indicates the presence of dark energy as a large part of the total energy content of the universe as we will discuss in Chap. 15. The observed acceleration is quite important as it is one of the bases of the presently favored cosmological model, the Λ CDM or LCDM model, where CDM stands for cold dark matter and Λ or L refers to the cosmological constant λ .

B. *Black body radiation fills the Universe.* This radiation, called the cosmic microwave background or CMB, is observed directly with satellite and ground-based microwave and infrared detectors and fits the Planck black body spectrum for a temperature 2.75 K extremely well. It is also very isotropic in direction. The standard interpretation is that it is the remnant of the thermal radiation produced by the very hot big bang explosion, now greatly cooled by the expansion of the universe. The CMB existence and nature leave little doubt that the general big bang and expanding universe scenario are correct (Kirshner 2004). Indeed we can think of the CMB pattern on the sky as a photo of the big bang.

Theoretical models of the early universe predict the detailed spectrum of the CMB, that is the very small deviations from a perfect isotropic black body spectrum, of order 10^{-5} . For example the LCDM theory of the early universe with inflation explains some detailed properties of the CMB spectrum, so the spectrum has become a standard and useful tool for testing the LCDM model and the early universe (Chalinor 2012; NASA 2019).

For example the CMB spectrum has become an important tool to measure the Hubble constant independently of the classic Hubble diagram technique discussed above (Planck 2018; NASA 2019). The result of the CMB measurements and theoretical analysis is a value of about $H_0 = 67$ (km/s)/Mpc. It is gratifying that this value is in rough agreement with the classic technique of Hubble that we mentioned

above, but the error bars of the separate techniques do not overlap, so the agreement is not good enough to satisfy many cosmologists (Reiss 2019; Planck 2018). Some cosmologists refer to the situation as a “tension” rather than an outright disagreement (Crane 2019; Freedman 2019). Appendix 1 has more information on estimates for the Hubble constant and its uncertainty. Because of the tension we will here use only the rough conservative estimate $H_0 = 70 \pm 5$ (km/s)/Mpc for our pedagogical purposes.

Finally we note that the Hubble constant can also be determined a number of other ways. One example is to use gravitational lensing to determine distances (Schutz 2009; Chen 2019); another is to use gravitational wave data to determine both distance and velocity of black hole and neutron star sources, which we alluded to in Chap. 11 (Holz 2018). See Appendix 1 for more information on such measurements (Schutz 1986; Holz 2005).

- C. *The distribution of visible matter on the largest scale is approximately homogeneous and isotropic.* On a cosmological scale the distribution of clusters of galaxies, the most visible matter of the universe, appears to be homogeneous and isotropic—that is approximately the same in all directions and uniform in space. On a smaller scale there is of course an obvious hierarchy of clustering—stars cluster into galaxies, galaxies form clusters, and so forth. On an intermediate scale, that is large compared to galaxies and small compared to the cosmological scale, the universe has sheets and filaments of galaxies with large voids between them. It has been compared to a foam of liquid, for example the head of foam on a glass of beer.

For our theoretical study we use three fundamental assumptions, which are largely based on the above observations. Like all theoretical assumptions they should not be treated as absolute, but subject to further experiments and observations.

1. *Gravity, as described by general relativity, dominates the universe.* No other forces appear to be relevant on a cosmological scale. For example electric and magnetic fields are important on a stellar scale and for clusters of stars, but become less important on a galactic and cosmological scale. Thus the gravitational field equations of general relativity are assumed to describe the universe.
2. *The cosmological material can be treated as a perfect cosmic fluid.* The many billions of galaxies that now make up the visible universe behave as a low-pressure perfect fluid. For earlier times the universe was undoubtedly dominated by radiation and hot gases, which also behaved as a perfect fluid with high pressure. The invisible dark matter that appears to be a fundamental component of the present universe apparently also behaves like a low-pressure perfect fluid. Finally the dark energy behaves like a perfect fluid with pressure equal to the negative of the constant energy density. For the very earliest times assumptions about the nature of the cosmic fluid vary widely, as we will later discuss.
3. *On the cosmological scale the geometry of the universe is approximately homogeneous and isotropic.* Because of the isotropy and homogeneity of the visible galaxies and the observed isotropy of the black body radiation we assume that the

metric of the universe on the largest cosmological scale is also homogeneous and isotropic. This symmetry is a very strong constraint and will allow us to simplify the mathematical problem and make it tractable. This assumption is basic and powerful: it almost completely determines the general form of the metric as we will show in the next section.

13.2 The Cosmological FLRW Metric

The observations and related assumptions in the preceding section place a rigid constraint on the cosmological metric: it must represent a 3-dimensional space that is homogeneous and isotropic—the same everywhere and the same when viewed in any direction. We can obtain the general form of the metric of this 3-space if we first consider the analogous 2-space problem, which is intuitive and can be visualized. Time of course must be added to the space dimensions to give the final metric in spacetime.

In two dimensions there are two spaces that immediately come to mind that are homogeneous and isotropic: the Euclidean plane and the surface of a sphere. The metric on the surface of a sphere with radius R was obtained in Chap. 4, but we will repeat it here. In cylindrical coordinates (ρ, φ, z) the Euclidean 3-space metric and the constraint equation for a sphere are

$$d\ell^2 = d\rho^2 + \rho^2 d\varphi^2 + dz^2, \quad \rho^2 + z^2 = R^2. \quad (13.2)$$

In this section we will use $d\ell^2$ for the spatial line elements, and reserve ds^2 for the cosmological metric of space-time. We calculate dz from the constraint equation and substitute it into the 3-space metric to get the 2-space metric for the surface of the sphere

$$d\ell^2 = \frac{d\rho^2}{1 - \rho^2/R^2} + \rho^2 d\varphi^2. \quad (13.3)$$

Figure 13.3 shows the polar coordinates on the sphere. For $\rho \ll R$, near the north pole, the coordinates are like plane polar coordinates. The radial coordinate ρ cannot

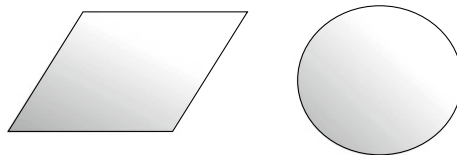


Fig. 13.3 The 2-spaces that are obviously homogeneous and isotropic—the plane and the sphere. They have analogs in three dimensions, where they are called the hyperplane and the hypersphere

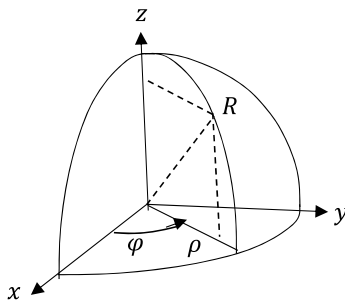


Fig. 13.4 Definition of polar coordinates on the surface of a sphere. One octant is shown

exceed R and for $\rho = R$ the g_{11} metric component is singular. Clearly only half the sphere is covered by these coordinates.

Now we use a little trick and introduce a curvature parameter k defined as $k = 1$ for a sphere and $k = 0$ for a plane. Then the metric for both the plane and the sphere in Fig. 13.4 may be written in one form,

$$d\ell^2 = \frac{d\rho^2}{1-k\rho^2/R^2} + \rho^2 d\varphi^2, \quad \begin{array}{ll} k = 0, & \text{plane} \\ k = 1, & \text{sphere} \\ k = -1, & \text{pseudosphere} \end{array} \quad (13.4)$$

But notice that we have added $k = -1$ to the list of surfaces in (13.4). It is an interesting fact that the metric (13.4) with $k = -1$ also represents a homogeneous and isotropic space, but one which cannot be visualized as a surface in a Euclidean 3-space like Fig. 13.3. It is called a *pseudosphere*, and it is possible to study its mathematical properties despite the fact that we cannot visualize it.

The three 2-spaces in (13.4) are homogeneous and isotropic, although this may not be readily apparent for $k = -1$. Below we will generalize (13.4) to three dimensions and later add time to get the cosmological metric.

Before going on to three dimensions it is useful to understand a little more about the geometric nature of these 2-spaces, especially the pseudosphere. Let us first calculate the ratio of the circumference C_s to the radius R_s of a small circle around the north pole in Fig. 13.4. To calculate the circumference we set the radial coordinate to a constant ρ and integrate the angular part of the metric. This gives

$$C_s = \int_0^{2\pi} \sqrt{g_{22}} d\varphi = \int_0^{2\pi} \rho d\varphi = 2\pi\rho. \quad (13.5a)$$

The radius R_s of the small circle is given by

$$R_s = \int_0^\rho \sqrt{g_{11}} d\rho = \int_0^\rho \frac{d\rho}{\sqrt{1 - k\rho^2/R^2}} \cong \rho \left(1 + k \frac{\rho^2}{6R^2} \right). \quad (13.5b)$$

Thus the ratio of circumference to radius is

$$\frac{C_s}{R_s} \cong 2\pi \left(1 - k \frac{\rho^2}{6R_s^2} \right). \quad (13.6)$$

The three 2-spaces are thus characterized by: $C_s/R_s < 2\pi$ for the sphere: $C_s/R_s = 2\pi$ for the plane: $C_s/R_s > 2\pi$ for the pseudosphere. Therefore we may think of the sphere as gotten from the plane by compressing space around any given point, and we may think of the pseudosphere as gotten from the plane by stretching space around any given point. This is illustrated in Fig. 13.5 for a little cap-shaped region on the sphere and a little saddle-shaped region on the pseudosphere.

For the spherical case we can perform such cutting and pasting to roughly construct a sphere. For the pseudosphere we cannot since there is not enough room in Euclidean 3-space!

Having obtained the 2-spaces described by (13.4) we have solved the problem of finding suitable homogeneous and isotropic spaces in two dimensions. Now we extend the analysis to three dimensions. In analogy with (13.2) we consider in coordinates (r, θ, φ, q) the Euclidean 4-space metric and the constraint equation defining a 3-sphere.

$$d\ell^2 = [dr^2 + r^2(d\theta^2 + \sin^2 \theta d\varphi^2)] + dq^2, \quad r^2 + q^2 = R^2. \quad (13.7)$$

Then by calculating dq from the constraint equation and substituting it in the metric we obtain the analog of (13.4) in three dimensions

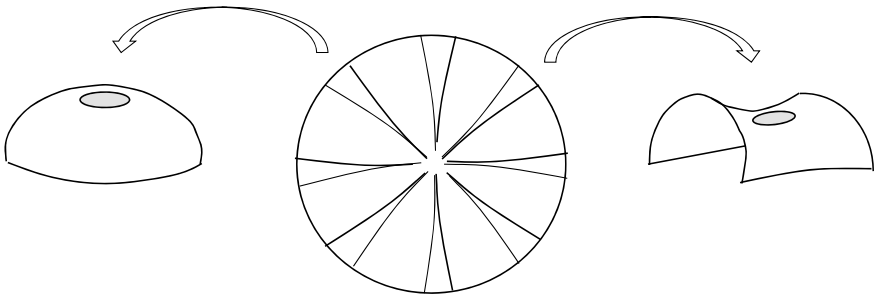


Fig. 13.5 From the disk we delete the wedge-shaped pieces, then glue the edges together and the resulting surface will fit on the spherical surface on the left. If we double the little wedge-shaped pieces the resulting surface will fit on the pseudospherical surface on the right. Of course we implicitly think of the limit of many little wedges covering finite regions of the surface

$$d\ell^2 = \frac{dr^2}{1 - kr^2/R^2} + r^2(d\theta^2 + \sin^2 \theta d\varphi^2), \quad \begin{array}{ll} k = 0, & \text{Euclidean 3-space} \\ k = 1, & \text{hypersphere} \\ k = -1, & \text{pseudohypersphere} \end{array} \quad (13.8)$$

All three of the spaces in (13.8) are homogeneous and isotropic, although this might not be obvious from the form of the metric. We can also write the metric in terms of a dimensionless radial coordinate defined by $w = r/R$ as

$$d\ell^2 = R^2 \left[\frac{dw^2}{1 - kw^2} + w^2(d\theta^2 + \sin^2 \theta d\varphi^2) \right]. \quad (13.9)$$

It is sometimes desirable to write the spatial metric (13.9) in a conformally flat form, that is proportional to flat Euclidean 3-space. To do this we introduce another dimensionless radial coordinate u by

$$w = \frac{u}{1 + ku^2/4}. \quad (13.10)$$

A little algebra gives the metric as

$$d\ell^2 = \frac{R^2}{(1 + ku^2/4)^2} [du^2 + u^2(d\theta^2 + \sin^2 \theta d\varphi^2)]. \quad (13.11)$$

This form is less commonly used but can be convenient for use with some coordinates, such as Cartesian (Adler 1975).

It is now straight-forward to get the cosmological metric. We think of the 3-space as expanding with time. This corresponds to the radius R in the 3-space metric increasing with time as $R(t)$. For the time component of the metric we choose a universal time coordinate which is the same as the proper time for a stationary observer. In terms of the dimensionless radial coordinates w we then write the cosmological metric in the form

$$ds^2 = c^2 dt^2 - R(t)^2 \left[\frac{dw^2}{1 - kw^2} + w^2(d\theta^2 + \sin^2 \theta d\varphi^2) \right]. \quad (13.12a)$$

The quantity R naturally has the dimension of a distance, so the spatial part of the metric has the dimension of a physical distance squared, as it must. The curvature parameter is $k = \pm 1, 0$ and dimensionless.

We will here adopt a more common convention for the metric, which is to take the radial coordinate r to have the dimension of a distance, and k to be a constant parameter with the dimension of an inverse distance squared. In this scheme the metric is

$$ds^2 = c^2 dt^2 - a(t)^2 \left[\frac{dr^2}{1 - kr^2} + r^2(d\theta^2 + \sin^2 \theta d\varphi^2) \right], \quad (13.12b)$$

and the function $a(t)$ is dimensionless. The function $a(t)$ is called the scale factor and is at the heart of cosmological theory. The choice (13.12b) has a nice advantage: since only the product of the scale factor times the bracket in (13.12b) has physical meaning we can consistently take $a(t)$ to be unity at the present cosmic time t_0 . Then the square bracket in (13.12b) is the square of the current physical distance to a nearby galaxy. We can think of the curvature parameter k as the inverse square of the maximum coordinate distance allowed for an observed galaxy.

The cosmological metric in either the form (13.12a) or (13.12b) is known as the Friedmann-Lemaitre-Robertson-Walker or FLRW metric, after its various discoverers (Schutz 2009). It is the basis of relativistic cosmological theory, and a primary task of theoretical cosmology is to determine the nature of the scale factor and the curvature parameter to compare with observations.

13.3 Consequences of the Metric

Before we go on to apply the field equations of general relativity to the cosmological problem we will show several cosmological consequences that follow from the FLRW metric alone, independent of the dynamics imposed by the field equations. The first consequence provides a remarkably simple picture of the motion of particles in the metric, of the galaxies in the cosmic fluid: we will show that they remain at fixed locations with respect to the coordinate system. This is of course an approximation that ignores local or peculiar motion of galaxies.

A galaxy is assumed to follow in general relativity a geodesic

$$\frac{d^2 x^\alpha}{ds^2} + \Gamma_{\beta\gamma}^\alpha \frac{dx^\beta}{ds} \frac{dx^\gamma}{ds} = 0. \quad (13.13)$$

For its spatial motion we take $\alpha = i$ and calculate the acceleration,

$$\frac{d^2 x^j}{ds^2} = -\Gamma_{\beta\gamma}^j \frac{dx^\beta}{ds} \frac{dx^\gamma}{ds}. \quad (13.14)$$

For a galaxy that is initially at rest in the coordinate system let us see what this acceleration is. For such a galaxy the spatial components of the 4-velocity are zero, so

$$\frac{d^2 x^j}{ds^2} = -\Gamma_{00}^j \frac{dx^0}{ds} \frac{dx^0}{ds}. \quad (13.15)$$

Thus we need only one type of connection,

$$\begin{aligned}\Gamma_{00}^j &= \frac{1}{2} g^{j\beta} (g_{0\beta,0} + g_{\beta 0,0} - g_{00,\beta}) = \frac{1}{2} g^{jj} (g_{0j,0} + g_{j0,0} - g_{00,j}) \\ &= -\frac{1}{2} g^{jj} g_{00,j}.\end{aligned}\quad (13.16)$$

The last step follows because the metric is diagonal. But the 0, 0 component of the FLRW metric is equal to 1, so the acceleration vanishes. Thus a galaxy initially at coordinate rest suffers no acceleration and remains at coordinate rest.

We emphasize that this means a galaxy stays at the same 3-space coordinate position, but not that it stays at rest physically; since the metric is time dependent it does indeed move physically. An often-used analogy with a 2-sphere is useful here; picture the galaxies as glued to the surface of a balloon on which a coordinate grid has been drawn with ink. As the balloon is inflated the galaxies move apart, even though they remain at the same place in the coordinate grid. Another analogy for the 3-plane is also apt. Picture an unbaked loaf of raisin bread which has been put in the oven to rise. As the bread rises and expands the raisins stay at the same position with respect to the dough, but because the dough expands they move apart physically. Because the galaxies remain at the same 3-space coordinate positions and move with the coordinate grid such 3-space coordinates are called co-moving coordinates. The simplicity of the galactic motion makes co-moving coordinates very useful, as we will find below.

The picture of the galaxies at coordinate rest is of course not exact; there is also individual motion of the galaxies in terms of the coordinates and in terms of physical motion. The individual motion is generally called the peculiar motion while the motion associated with the universal expansion is generally called the Hubble motion or Hubble flow. The peculiar motion is roughly of order 300 km/s.

A second consequence, a most important one, of the FLRW metric is the clear explanation it provides for the cosmological redshift and Hubble's law. Let us write the FLRW metric (13.12b) as

$$ds^2 = c^2 dt^2 - a(t)^2 d\sigma^2, \quad d\sigma^2 = \frac{dr^2}{1 - kr^2} + r^2(d\theta^2 + \sin^2\theta d\varphi^2). \quad (13.17)$$

The 3-space coordinate separation σ between any two co-moving galaxies is the integral of $d\sigma$, which remains constant in time as we have just seen. Consider then a galaxy emitting at time t_e a photon of light with period Δt_e , which then travels to us, the observers, arriving here at time t_o with period Δt_o . Figure 13.6 shows the scenario.

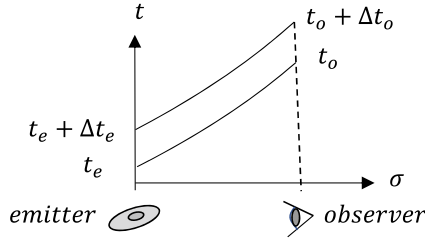


Fig. 13.6 The photon is emitted by a galaxy and travels to us while the 3-space coordinate separation σ of the galaxies remains fixed

During the photon's travel the 3-space co-moving coordinate distance σ remains fixed while the scale factor $a(t)$ changes. For the photon of light we recall the fundamental fact that it follows a path with a line element equal to zero, a null path,

$$ds^2 = c^2 dt^2 - a(t)^2 d\sigma^2 = 0, \quad \frac{cdt}{a(t)} = d\sigma. \quad (13.18)$$

We integrate this over the travel time of the photon from t_e to t_o to give σ . We also integrate from $t_e + \Delta t_e$ to $t_o + \Delta t_o$ to give the same σ since the galaxies are co-moving, stationary in the coordinate system; thus

$$\int_{t_e}^{t_o} \frac{cdt}{a(t)} = \int_{t_e + \Delta t_e}^{t_o + \Delta t_o} \frac{cdt}{a(t)} = \sigma. \quad (13.19)$$

We now take the difference between these two equal integrals, and find approximately

$$\begin{aligned} \int_{t_e + \Delta t_e}^{t_o + \Delta t_o} \frac{cdt}{a(t)} - \int_{t_e}^{t_o} \frac{cdt}{a(t)} &= \int_{t_o}^{t_o + \Delta t_o} \frac{cdt}{a(t)} - \int_{t_e}^{t_e + \Delta t_e} \frac{cdt}{a(t)} = \frac{c \Delta t_o}{a(t_o)} - \frac{c \Delta t_e}{a(t_e)} = 0, \\ \frac{\Delta t_o}{\Delta t_e} &= \frac{a(t_o)}{a(t_e)}. \end{aligned} \quad (13.20)$$

Thus we see that the period of the light increases as the universe expands from $a(t_e)$ to $a(t_o)$. This is called the cosmological redshift. In terms of wavelength or frequency of the light it may equivalently be written as

$$\frac{\lambda_o}{\lambda_e} = \frac{\nu_e}{\nu_o} = \frac{a(t_o)}{a(t_e)}. \quad (13.21)$$

This beautiful result tells us that as the photon travels its wavelength stretches in proportion to the scale factor of the universe. It is a very easy way to remember the cosmological redshift relation.

Let us proceed to get the Hubble law (13.1) for cosmologically nearby galaxies. Astronomers define a redshift parameter z as the fractional wavelength shift; according to (13.21) z is related to the scale factor by

$$z = \frac{\Delta\lambda}{\lambda} = \frac{\lambda_o - \lambda_e}{\lambda_e} = \frac{\lambda_o}{\lambda_e} - 1 = \frac{a(t_o)}{a(t_e)} - 1. \quad (13.22)$$

A cosmologically nearby galaxy then has $z = 0$.

A receding galaxy appears to be moving away from us at a velocity given by the Doppler shift, so

$$\frac{v}{c} = \frac{\Delta\nu}{\nu} = \frac{\Delta\lambda}{\lambda} = z = \frac{a(t_o) - a(t_e)}{a(t_e)}. \quad (13.23)$$

For a cosmologically nearby galaxy the difference between the time of emission and observation is small so we may expand $a(t)$ to obtain

$$\frac{v}{c} = \frac{a'(t_o)}{a(t_o)}(t_o - t_e), \quad a' = \frac{da}{dt}. \quad (13.24)$$

The distance to such a nearby galaxy is approximately $L = c(t_o - t_e)$, so that the velocity of recession is

$$v = \frac{a'(t_o)}{a(t_o)}L. \quad (13.25)$$

We have thus derived Hubble's law (13.1) and identified the Hubble constant in (13.25). Following current use we define a Hubble function H whose current value is the Hubble constant,

$$H(t) \equiv \frac{a'(t)}{a(t)}, \quad H_0 = H(t_0) = \frac{a'(t_0)}{a(t_0)}. \quad (13.26)$$

Note that some authors refer to $H(t)$ as the Hubble parameter, which somewhat obscures its nature as a function of time.

To summarize, the FLRW metric with an increasing scale factor implies a cosmological redshift and the Hubble law, with the Hubble constant simply related to the present scale factor. This gives a very useful observational constraint on the scale factor.

But we need to continue the analysis of the redshift distance relation to higher order since observations now justify more accuracy, as we indicated at the beginning of this chapter. The analysis will be quite useful in later chapters; the algebra is somewhat tedious but straight-forward. We first expand the redshift relation (13.23) to second order in travel time,

$$z = \frac{a(t_o)}{a(t_e)} - 1 = \frac{a'(t_o)}{a(t_o)}(t_o - t_e) + \left[\left(\frac{a'(t_o)}{a(t_o)} \right)^2 - \frac{a''(t_o)}{2a(t_o)} \right] (t_e - t_o)^2. \quad (13.27)$$

To write this in a prettier way we define a dimensionless deceleration function $q(t)$ proportional to the deceleration of the scale factor, and call its present value q_0 ,

$$q(t) = -\frac{a''(t)a(t)}{a'^2(t)} = \frac{a''(t)}{H^2(t)a(t)}, \quad q_0 = -\frac{a''(t_o)}{H_0^2 a(t_o)}. \quad (13.28)$$

(Recall that we may take the present value of the scale factor to be unity as we have discussed.) In terms of the Hubble constant H_0 and the deceleration constant q_0 the redshift expression (13.27) becomes somewhat prettier

$$z = H_0(t_o - t_e) + H_0^2 \left(1 + \frac{q_0}{2} \right) (t_o - t_e)^2. \quad (13.29)$$

It remains to write the redshift z in terms of the galactic distance rather than the travel time. For this we need to calculate the galactic distance as an expansion in the light travel time. The coordinate distance to the galaxy is given by integrating (13.18), so we obtain by expansion

$$\begin{aligned} \sigma &= \int_{t_e}^{t_o} \frac{cdt}{a(t)} = \int_{t_e}^{t_o} \frac{cdt}{a(t_o) + a'(t_o)(t - t_o)} = \frac{c}{a(t_o)} \int_{t_e}^{t_o} dt \left[1 - \frac{a'(t_o)}{a(t_o)}(t - t_o) \right] \\ &= \frac{c}{a(t_o)}(t_o - t_e) + \frac{cH_0}{2a(t_o)}(t_o - t_e)^2. \end{aligned} \quad (13.30)$$

The physical distance L is thus

$$L = a(t_o)\sigma = c(t_o - t_e) + cH_0(t_o - t_e)^2/2. \quad (13.31)$$

Inverting this to second order we get the travel time in terms of the physical distance,

$$(t_o - t_e) = \frac{L}{c} - \frac{H_0}{2c^2} L^2. \quad (13.32)$$

Finally we substitute this into (13.29) to get the redshift distance relation to second order,

$$z = \left(\frac{H_0 L}{c} \right) + \frac{1}{2} (1 + q_0) \left(\frac{H_0 L}{c} \right)^2, \quad H_0 = \frac{a'(t_0)}{a(t_0)}, \quad q_0 = -\frac{a''(t_0)}{H_0^2 a(t_0)}. \quad (13.33)$$

This is just the Hubble law with $v/c = z$ plus a second order correction in the distance.

When it was first introduced the quantity q_0 was called the deceleration parameter because it was expected to be positive, corresponding to the deceleration of the scale factor expected for a matter dominated universe; however as we have noted above nature does not work that way and it turns out that q_0 is negative, and the universe accelerates as we will discuss further.

In practice the task of observational cosmology is to fit the data for galaxies or supernovas to (13.33) to obtain values for the Hubble constant H_0 and q_0 , which we will discuss below.

For some specific metrics the rather tedious expansion analysis leading to the approximate redshift distance relation (13.33) can be replaced by an exact calculation; for example it can be done exactly for de Sitter space, as discussed below in Sect. 13.4 and Exercise 13.10.

As our last illustration of the use of the FLRW metric we will study the physical distance to a distant galaxy. Suppose we place ourselves at the center of the coordinate system, $r = 0$. How far away is a galaxy at coordinate radius r ? The relation between the radial coordinate distance and physical distance for the diagonal FLRW metric is

$$\ell = \int_0^r \sqrt{|g_{11}|} dr' = a(t_0) \int_0^r \frac{dr'}{\sqrt{1 - kr'^2}}. \quad (13.34)$$

This may be integrated exactly to give

$$\begin{aligned} \ell &= a(t_0) \arcsin(r\sqrt{k}) \frac{1}{\sqrt{k}} & \text{for } k > 0 \\ \ell &= a(t_0) r & \text{for } k = 0 \\ \ell &= a(t_0) \operatorname{arcsinh}(r\sqrt{|k|}) \frac{1}{\sqrt{|k|}} & \text{for } k < 0 \end{aligned} \quad (13.35)$$

Alternatively, for relatively nearby galaxies, we may approximate the distance as

$$\ell \cong a(t_0) \int_0^r \left(1 + \frac{kr'^2}{2} \right) dr' = a(t_0) r (1 + kr^2/6). \quad (13.36)$$

The approximate form is rather elegant. Note that we may again use $a(t_0) = 1$ if desired. Notice also from (13.35) that for positive k we may interpret $1/\sqrt{k}$ as the maximum distance of a galaxy from us.

Equation (13.36) is also useful for illustrating an interesting geometric concept. We can carry out in three dimensions the same calculation that led to (13.6) in 2 dimensions, that is the ratio of the circumference C_s to the radius R_s of a small circle as illustrated in Fig. 13.5. We obtain for the present three dimensional case

$$\frac{C_s}{R_s} \cong 2\pi \left(1 + k \frac{r^2}{6} \right). \quad (13.37)$$

This is the same result as (13.6) although in slightly different notation: in (13.6) the curvature parameter k is dimensionless and in (13.37) it is an inverse distance squared. This clearly tells us that, loosely speaking, there is “too little space” around a given point in the hypersphere and “too much space” in the pseudo-hypersphere.

It is also easy to relate physical distances to the radial coordinate u used in (13.10) and (13.11), which we leave to the reader in Exercise 13.16.

13.4 De Sitter Space

It is amusing to consider a simple cosmological model based only on ad hoc considerations and not on the gravitational field equations. Without physical justification we suppose that the Hubble constant is truly constant, $H(t) = H_0$. Then we may integrate (13.26) to obtain the scale factor at any time, which we write as

$$a(t) = a(t_0)e^{H_0(t-t_0)}. \quad (13.38)$$

This is known as the de Sitter model universe. It is a perpetual universe: it has always existed and always will exist, expanding exponentially forever. It was extremely small in the distant past, but never had zero size. If we also assume $k = 0$ the metric is quite simple

$$ds^2 = c^2 dt^2 - a(t_0)^2 e^{2H_0(t-t_0)} [dr^2 + r^2(d\theta^2 + \sin^2\theta d\varphi^2)]. \quad (13.39)$$

The space-time is called de Sitter space. We will return to de Sitter space in later chapters. It is intimately related to the actual universe in the distant future, and also has features in common with the very early inflationary universe. Later we will also use the field equations and consider spatially non-flat versions with $k \neq 0$.

Appendix 1: Measured Values for the Hubble Constant

The original Hubble method to obtain H_0 requires that we measure the recession velocity and the distance to a number of galaxies or other cosmological sources and then fit a plot of the two to (13.33). The recession velocity is relatively easy to measure using the Doppler shift of the light. However measuring the distance is not so simple. It is typically done using a so-called distance ladder; we first use parallax to measure the distance to nearby sources such as Cepheid variable stars, whose period varies in a known way with their intrinsic brightness or absolute luminosity, making them “standard candles.” This makes both the apparent luminosity and the absolute luminosity of the standard candles measurable, and from that the distance can be calculated; the next step on the ladder is to use the Cepheid variable stars to measure the distance and brightness of type Ia supernovae whose spectra can be correlated with their absolute luminosity so that they also serve as standard candles. The result is that we can determine the absolute luminosity of yet more distant supernovae by observing their time spectra, and the combination of apparent and absolute luminosity then gives their distance. The basic idea is further discussed in Chap. 12 of Schutz (2009) and a recent detailed application is in Reiss (2019).

Another approach to obtaining H_0 is to use red giant stars as standard candles (Freedman 2019). The gravitational lensing of galaxies can also be used to measure the distance to a source; the basic idea is discussed in Schutz (2009) and the application to measuring H_0 in Chen (2019). Also see Exercise 13.17 on lensing. Finally, the collision and merger of black holes and neutron stars provide “standard sirens” that allow a measurement of H_0 from observations of the gravitational waves they emit (Holz 2018, 2005; Schutz 1986).

The value for the Hubble constant obtained from the ladder approach is about $H_0 = 74$ (km/s)/Mpc. This is widely called the “local” value for obvious reasons. Figure 13.7 and Table 13.1 show specific values and error estimates.

The CMB spectrum provides a conceptually different approach to measuring H_0 . We can use theory, such as the LCDM model, to estimate the scale factor when the CMB was emitted in terms of cosmological parameters such as H_0 and some present-day density ratios which we will discuss later in Chap. 14 (see specifically (14.19)).

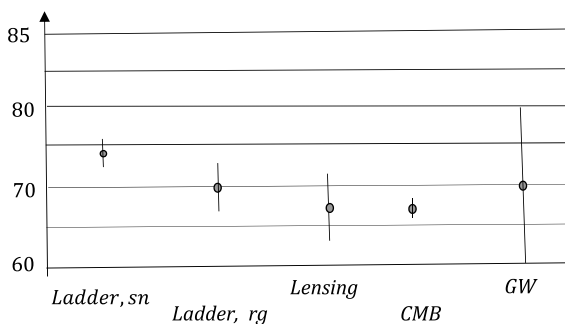


Fig. 13.7 Values of the Hubble constant in (km/s)/Mpc; sn denotes supernovae, rg denotes red giants, and GW denotes gravitational waves

Table 13.1 Values for the hubble constant

Method	Value	Error (approx.)	Reference
Ladder, <i>sn</i>	74.03	1.4	Reiss (2019)
Ladder, <i>rg</i>	69.8	2.5	Freedman (2019)
Lensing	67.4	4.1	Birrer (2020)
CMB	67.36	0.5	Planck (2018)
GW	70	10	Holz (2019)

The theoretical CMB spectrum obviously depends on the scale factor at the time of emission as well as other physical properties such as the energy density over time of the radiation, and the velocity of sound and standing waves in the cosmic fluid (Knox 2019). (We will say more in Sect. 17.4.) By comparing the theoretical CMB spectrum with the observed spectrum we can thus make a best fit that gives values for the various cosmological parameters, in particular H_0 . The value obtained in this way is about $H_0 = 67$ (km/s)/Mpc (Chaliner 2012; Planck 2018; NASA 2019)

Figure 13.7, along with Table 13.1, shows some of the interesting observational results. It includes ladder method results using supernovae, red giants and gravitational lensing (Reiss 2019; Freedman 2019; Chen 2019). The CMB result using the Planck satellite is shown, as well as the standard siren result based on gravitational wave observations by LIGO (Planck 2018; Holz 2018). The values are clearly not in violent disagreement, but the error estimates do not overlap; this causes concern for many cosmologists (Crane 2019; Reiss 2019; Planck 2018).

The supernova ladder and CMB values differ by roughly 4 times more than the observers error estimates (Reiss 2019). The red giant value lies between the supernova value and the CMB value. The lensing value depends on the method of data analysis. From this it appears that either the observers are overly optimistic concerning their error estimates, or there is a problem with the basic theory and we must go beyond the LCDM model. One example is that the number of neutrino types assumed may be incorrect. Another is that the exponent for the dark energy term in (14.19) is not correct. There are many possibilities (Knox 2019).

A historical note is in order. As we noted in Sect. 13.1, for decades the value of H_0 was unknown to about a factor of 2; the values favored by different observer groups were about 50 and 100. No basic theoretical ideas emerged from this disagreement, and it was resolved by later observations. For our pedagogical purposes we have used the rough error estimate of about 5 (km/s)/Mp that pessimistically encompasses all the individual error estimates.

Exercises

- 13.1 Take the radius of the observable universe to be roughly 10 billion light years, and the separation between galaxies to be roughly 1 million light years. Very roughly, how many galaxies are there in the observable universe? How many stars? How many planets?

- 13.2 Look up the nature and size of the great walls and voids in the matter distribution of the universe. Compare to the cosmological scale of about 10 billion light years.
- 13.3 In Euclidean plane geometry parallel lines never meet, and only parallel lines never meet. What is the analog of this statement for the surface of a sphere and a pseudosphere?
- 13.4 Calculate the Riemann scalar for the surface of a sphere and a plane and a pseudosphere. How is it related to the parameters R and k .
- 13.5 Consider the metric on a sphere in (13.4) using cylindrical coordinates. Transform to the usual spherical coordinates using $\rho = R \sin \theta$ and get the more standard form

$$d\ell^2 = R^2(d\theta^2 + \sin^2 \theta d\varphi^2).$$

- 13.6 Let us do the analog of Exercise 13.5 in 3 dimensions. Consider the 3-sphere metric in (13.8). Introduce a hyperspherical angle ψ defined by $r = R \sin \psi$, and show that the metric becomes

$$d\ell^2 = R^2[d\psi^2 + \sin^2 \psi (d\theta^2 + \sin^2 \theta d\varphi^2)].$$

This is useful for many geometric calculations. Can you construct the metric for a 4-sphere in a similar way? Do you see the pattern?

- 13.7 Calculate the circumference of the hyperspherical universe.
- 13.8 Calculate the total volume of the hyperspherical universe.
- 13.9 For the pseudo-hypersphere, $k = -1$, there is a singularity in the metric (13.11) at $u = 2$. Discuss briefly the behavior of the metric and the space there.
- 13.10 Consider a de Sitter universe with curvature parameter $k = 0$ and Hubble constant H_0 . If a galaxy has a redshift of z how far away is it? What numbers do you get for a Hubble time of 14 billion years and a redshift of $z = 2$? Is there any upper limit to the redshift and the distance of the galaxy?
- 13.11 In the text we discussed the cosmological metric as the 3-dimensional generalization of the plane and sphere and pseudosphere. It is possible to also construct other 2-spaces that are homogeneous and isotropic but topologically more complex. For example, consider a flat square torus; it is constructed by identifying the opposite sides of an ordinary square. Simply glue them together! This space is clearly locally homogeneous and finite. Can you do an analogous construction for the surface of a sphere and pseudosphere?
- 13.12 Suppose the space part of the metric of the universe is in fact a *cubic torus*, the three dimensional analog of the square torus in Exercise 13.11. What observations could you make to test this idea? Can you think of any problems with such a theoretical speculation?
- 13.13 In the text we related the radial marker r used in (13.12b) to the physical distance ℓ . Do the same for the radial marker w used in (13.12a).

- 13.14 Why is it that we can study the mathematics of a pseudosphere but cannot actually construct one in Euclidian 3-space? Is there any logical inconsistency here? What of the 3-space for $k = -1$?
- 13.15 There is a hybrid convention possible regarding the FLRW metric. Choose L to be a convenient constant distance parameter, and write the metric as

$$ds^2 = c^2 dt^2 - a(t)^2 L^2 \left[\frac{dw^2}{1 - kw^2} + w^2 (d\theta^2 + \sin^2 \theta d\varphi^2) \right].$$

Then a can be dimensionless, w also dimensionless, and the parameter k can be ± 1 or 0. Show that this convention is consistent.

- 13.16 Repeat the calculations in (13.35) and (13.36) for distances but using the coordinate u in (13.10) and (13.11).
- 13.17 Use a reference on gravitational lensing such as Schneider (1992) and work out how such lensing can give the distance to a source.

Chapter 14

The Dynamical Equations of Cosmology



Abstract The Einstein equations applied to the FLRW metric give basic dynamical equations for cosmology, specifically for the scale factor. The dynamical equations depend on the physical properties of the constituents of the cosmic fluid, which we take to be vacuum or dark energy, cold matter, radiation, and an effective curvature. Together with the behavior of the constituents the dynamical equations lead to what we here call the Friedmann master equation for the scale factor of the universe; it is remarkably useful.

14.1 The Einstein Field Equations for Cosmology

In the preceding chapters we have set up the infrastructure of cosmology, and now we need to add dynamics via the Einstein field equations applied to the FLRW metric and the perfect fluid energy-momentum tensor (Adler 1975; Misner 1973; Peebles 1993). We repeat here from the last chapter the FLRW metric, the cosmic perfect fluid energy-momentum tensor, and the field equations in mixed index form.

$$ds^2 = c^2 dt^2 - a(t)^2 \left[\frac{dr^2}{1 - kr^2} + r^2 (d\theta^2 + \sin^2 \theta d\varphi^2) \right], \quad (14.1a)$$

$$T^\mu{}_\nu = \rho u^\mu u_\nu + p(u^\mu u_\nu - g^\mu{}_\nu), \quad (14.1b)$$

$$G^\mu{}_\nu + \Lambda g^\mu{}_\nu = CT^\mu{}_\nu = -(8\pi G/c^4)T^\mu{}_\nu. \quad (14.1c)$$

We use the form of the FLRW metric (13.12b) with dimensionless scale factor $a(t)$, radial coordinate r with the dimension of a distance, and an energy-momentum tensor with the dimensions of energy density; here ρ and p both have the dimensions of energy density or mass density times c^2 . The velocity u^μ is dimensionless. Either the mixed index form of the tensors or the lower index forms are convenient to use. Recall that the mixed metric tensor $g^\mu{}_\nu$ is the Kronecker $\delta^\mu{}_\nu$.

We take the cosmic fluid to be co-moving as we discussed in Chap. 13. From the definition of the 4-velocity and the FLRW metric we then obtain the 4-velocity of the fluid

$$u^\mu = \frac{dx^\mu}{ds} = (1, 0, 0, 0), \quad u_\beta = g_{\beta\alpha}u^\alpha = (1, 0, 0, 0). \quad (14.2)$$

From this the energy-momentum tensor on the right side of the field equations is

$$T^\mu{}_\nu = (\rho + p)u^\mu u_\nu - p\delta^\mu_\nu = \begin{pmatrix} \rho & 0 & 0 & 0 \\ 0 & -p & 0 & 0 \\ 0 & 0 & -p & 0 \\ 0 & 0 & 0 & -p \end{pmatrix}. \quad (14.3)$$

This is a wonderfully simple form for the right side of the field equations.

To get the geometric left side of the field equations we need the Einstein tensor. This involves slightly tedious but straight-forward algebra, so we have relegated it to Appendix 1, and Exercises 14.1 and 14.2. The diagonal components of the Einstein tensor are the following simple functions of the scale factor $a(t)$ and its derivatives,

$$\begin{aligned} G^0{}_0 &= -3 \left[\frac{k}{a^2} + \frac{a'^2}{c^2 a^2} \right], \quad a' \equiv \frac{da}{dt}, \\ G^1{}_1 = G^2{}_2 = G^3{}_3 &= - \left[\frac{k}{a^2} + \frac{a'^2}{a^2 c^2} + \frac{2a''}{ac^2} \right], \end{aligned} \quad (14.4)$$

and the off-diagonal components are identically zero. We substitute these and the energy-momentum tensor (14.3) into the field equations, to obtain

$$- \left(\frac{8\pi G}{c^4} \right) \rho = \Lambda - 3 \left[\frac{k}{a^2} + \frac{a'^2}{c^2 a^2} \right], \quad (14.5a)$$

$$\left(\frac{8\pi G}{c^4} \right) p = \Lambda - \left[\frac{k}{a^2} + \frac{a'^2}{a^2 c^2} + \frac{2a''}{ac^2} \right]. \quad (14.5b)$$

These Einstein field equations are the basis of standard cosmological theory. They are widely referred to as the Friedmann equations. A prime task of theoretical cosmology is to solve them for the scale factor. To do this we need to choose an appropriate source density and pressure or some relation between the two.

Three prime tasks of observational cosmology are to determine the curvature parameter k and the value of the cosmological constant Λ in (13.5), and of course compare observations with theory.

14.2 Critical Density and the Shape of the Universe

In order to put the cosmological equations (14.5) in a convenient and elegant form we first study the density of the universe; we will see that there is a critical density that determines the sign of the curvature parameter k . We rewrite the first cosmological equation (14.5a) with the curvature parameter k on the right side

$$\left[\left(\frac{8\pi G}{c^4} \right) \rho + \Lambda \right] - \frac{3}{c^2} \left(\frac{a'}{a} \right)^2 = \frac{3k}{a^2}. \quad (14.6)$$

Using the energy density of the vacuum as defined in (12.22) we may write the relation (14.6) in terms of densities,

$$(\rho + \rho_V) - \rho_{\text{crit}} = \left(\frac{3c^4}{8\pi G a^2} \right) k, \quad \rho_V \equiv \frac{c^4}{8\pi G} \Lambda, \quad \rho_{\text{crit}} \equiv \frac{3c^2 H^2}{8\pi G}. \quad (14.7)$$

This is an interesting equation; it tells us that if the total energy density of the universe $(\rho + \rho_V)$ is greater than the critical density ρ_{crit} as defined in (14.7) then the value of the curvature parameter k must be positive, if it is equal to the critical density then the curvature parameter must be zero, and if it is less than the critical density then the curvature parameter must be negative. This determines the geometric nature or shape of the universe, whether it is a 3-sphere or a 3-plane or a 3-pseudosphere. Moreover the critical density depends on the Hubble function, which is directly measurable at the present time. Because of this there is strong motivation to measure accurately the present energy density of the universe and the Hubble constant.

Equation (14.7) is often written with the densities expressed as fractions of the critical density; in terms of these fractional densities it becomes

$$\Omega + \Omega_V + \Omega_k = 1, \quad \Omega = \frac{\rho}{\rho_{\text{crit}}}, \quad \Omega_V = \frac{\rho_V}{\rho_{\text{crit}}} = \frac{\Lambda c^2}{3H^2},$$

$$\Omega_k = - \left(\frac{c^2}{H^2 a^2} \right) k. \quad (14.8)$$

The Ω are dimensionless density ratios: Ω_V denotes the effective vacuum density due to the cosmological constant, or dark energy density; Ω_k denotes an effective “curvature density ratio” as defined in (14.8), and is introduced mainly for notational convenience. We may think of the sum of the three density ratios on the left side of (14.8) as a sort of “total density ratio” that must equal unity by virtue of the field equation (14.6). This relation is important because of the link it provides with observation and for its use in obtaining an important equation of Friedmann for calculating the scale factor, which we will obtain in Sect. 14.5.

In the next section we will briefly discuss the observational values of the density ratios at the present time. In Sect. 16.3 we will return to the question of the shape of the universe.

14.3 Observed Dark Matter and Dark Energy Densities

The present observational value of the Hubble constant, as discussed in Chap. 13, is $H_0 = 70 \pm 5 \text{ (km/s)/Mpc}$, so the Hubble time is $14.0 \times 10^9 \text{ year}$. This gives a critical energy density of $8.28 \times 10^{-10} \text{ J/m}^3$ or a critical mass density of about $0.92 \times 10^{-26} \text{ kg/m}^3$, which is roughly 1 hydrogen atom per cubic meter on a global average. The measured mass density due to visible galaxies and other matter is of order 10^{-28} kg/m^3 , several orders of magnitude less than the critical value; however this does not mean that the curvature parameter is negative, since there may well be other significant matter present in the universe that is not visible. The visible or ordinary matter only gives a lower bound. Indeed the study of stars in galaxies and of galaxies in galaxy clusters indicates the presence of unseen *dark matter* producing a gravitational field; the amount of this dark matter appears to be quite significant (Rubin 1995, 1997).

Observations of stars and gas on the edges of spiral galaxies indicate that the matter orbits the center of the galaxy with a velocity v that is approximately constant, independent of the distance r from the center. This is quite surprising: most of the visible matter in a galaxy is in a small central bulge, so one expects the velocity to fall off like the square root of the distance from the center; this is easy to see by considering circular Newtonian orbits as noted in Exercises 14.3 and 14.4.

It thus appears that there must be mass present that is not visible and not concentrated at the center of the galaxy. If we assume that such dark matter is distributed roughly spherically about the galactic center with density $\rho(r)$ then the constant orbital velocity tells us that the density should be roughly proportional to $1/r^2$. This galactic halo of dark matter must extend well beyond the visible parts of the galaxy, as indicated in Fig. 14.1. Such a density profile is characteristic of an isothermal gas,

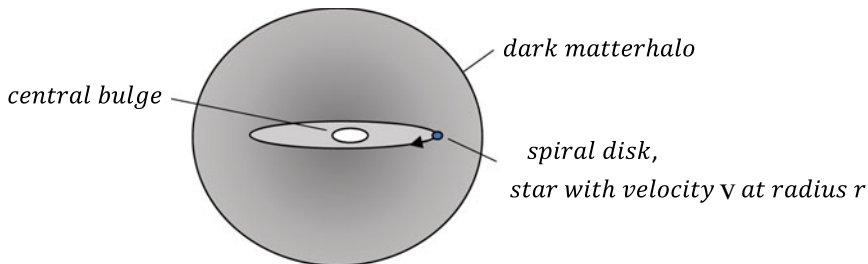


Fig. 14.1 General shape of a galaxy with a bright central bulge and disk imbedded in a halo of dark matter

although this correspondence should not be taken to be definitive since it implies an infinite total mass for the halo. See Exercise 14.5.

The existence of dark matter was first suggested by Zwicky in the 1930s; Zwicky studied clusters of galaxies and from their random velocities estimated their mass. See Exercise 14.6 (Zwicky 1933; Wiki DM).

The physical nature of the dark matter is not evident from observation. It could be almost anything that does not interact with light. For example, some of the dark matter could be small nonluminous stars called brown dwarfs, or black holes, or substellar lumps of matter, or interstellar gas and dust etc. More exotic possibilities are heavy elementary particles not yet seen in the laboratory such as supersymmetric particles, nonzero mass neutrinos, speculative light elementary particles called axions etc. The field is open to speculation (Schutz 2009; Randall 2018).

For the purpose of cosmology one important characteristic of the unseen material is the ratio of pressure to energy density, what we have called w . For ordinary matter and heavy elementary particles the ratio is very small, whereas for very light elementary particles it is about $1/3$, characteristic of a hot gas. For light the ratio is exactly $1/3$. The first case is referred to as cold dark matter, and the latter case is referred to as hot dark matter. The currently prevalent opinion is that the dark matter is probably cold.

The search for the physical nature of dark matter using laboratory detectors has been long and intense and unsuccessful, and is still a very active field. At present we have only the evidence of astronomical observations (Randall 2018, Wiki DM).

Concerning the nature and density of dark energy we will say more about this in Chap. 15, but we note here that it is now generally believed to be the cosmological constant and constitutes a large fraction of the total density in the universe, as we will discuss below.

Concerning the magnitude of the various densities we note at this point that the favored values, consistent with present observations are that the dark energy density is about 70% of the critical density, the dark matter is about 25% of critical, visible matter is only about 5% and the total density is equal to the critical density; thus the universe looks to be spatially flat with $k = 0$. In terms of the present fractional densities $\Omega_V = 0.70$, $\Omega = \Omega_{\text{CDM}} = 0.25$, $\Omega_{\text{vis}} = 0.05$. Remarkably it thus appears that the dominant constituents in the cosmic fluid, dark matter and dark energy, are not directly visible and the fundamental nature of the dark matter is not understood. We have a reasonable understanding of only about 5% of the stuff of our universe. This could be taken as a demand for modesty concerning our success in our overall understanding of nature.

14.4 Evolution of Cosmic Fluid Constituents

This section will deal with the behavior of the constituents of the cosmic fluid, such as cold matter and radiation, during the evolution of the universe. We will first show how energy is conserved in the expansion of the universe. Then we will obtain the

dependence of the various constituent densities on the scale factor of the universe, which is a most important result.

We begin by subtracting (14.5a) from (14.5b) to obtain

$$\left(\frac{4\pi G}{c^4}\right)(\rho + p) = \frac{k}{a^2} + \frac{a'^2}{a^2 c^2} - \frac{a''}{ac^2} = \frac{k}{a^2} - \frac{1}{c^2} \left(\frac{a'}{a}\right)'. \quad (14.9)$$

Notice that this relation does not depend on the cosmological constant. Next we differentiate (14.5a) with respect to time and find

$$\left(\frac{4\pi G}{c^4}\right)\rho' = \frac{-3ka'}{a^3} + \frac{3}{2c^2} \left(\frac{a'^2}{a^2}\right)' = -3\frac{a'}{a} \left[\frac{k}{a^2} - \frac{1}{c^2} \left(\frac{a'}{a}\right)'\right]. \quad (14.10)$$

Comparing (14.9) and (14.10) we see that

$$\rho' + 3(\rho + p)\frac{a'}{a} = 0. \quad (14.11)$$

Here the energy and pressure are the totals in the cosmic fluid. This first order relation will be useful for two purposes; the first is to demonstrate the conservation of energy during evolution of the universe, and the second is to show how individual densities, such as matter and radiation, behave as the universe expands, which is the main purpose in this section.

Equation (14.11) leads to an elegant statement of energy conservation in the cosmic fluid. We consider a small co-moving coordinate volume V_c , which of course remains constant during expansion, and the corresponding physical volume V , which increases with time as the universe expands; the two volumes are defined and related by

$$V_c = \int \frac{r^2 \sin \theta}{(1 - kr^2)} dr d\theta d\varphi = \text{const.}, \quad V = a^3 V_c. \quad (14.12)$$

Now we multiply (14.11) by $V = a^3 V_c$ to get

$$\rho' a^3 V_c + 3(\rho a^2 a' V_c + p a^2 a' V_c) = (\rho a^3 V_c)' + p(a^3 V_c)' = 0. \quad (14.13)$$

The last expressions in (14.13) have an important physical interpretation: the first term is the time derivative of the energy in the volume, and the second term is the pressure times the time derivative of the volume, so (14.13) may be expressed as

$$\frac{dE}{dt} + p \frac{dV}{dt} = 0. \quad (14.14)$$

This states that the change in the total energy in the co-moving volume is balanced by the work done on the volume by the pressure. It is the statement of cosmic energy conservation which we promised.

It is of prime importance to use (14.11) to analyze the separate evolution of the constituents of the cosmic fluid, in particular the matter and radiation energy densities. To do this we assume each constituent can be described by an effective linear equation of state, $p = w\rho$ where w is a constant parameter. Recall from Sect. 12.2 that according to the kinetic theory of gases the parameter is given by $w = (1/3)(v^2/c^2)$ where v^2 is the root-mean-square velocity of a gas molecule; as we saw, for the cold matter of the present universe $w = 0$ while for very hot gas or radiation $w = 1/3$. Equation (14.11) involves the total energy density and pressure in the cosmic fluid; we make the fundamental assumption that each constituent of the fluid separately obeys (14.11), which means that the constituents do not interact or at least do not interact strongly. This assumption is clearly reasonable for the cold matter and radiation in the present universe, but should be reconsidered for the earlier universe.

Substituting the linear equation of state $p = w\rho$ into (14.11) we obtain for each constituent

$$\frac{d\rho}{\rho} + 3(1+w)\frac{da}{a} = 0, \quad d[\log \rho + \log a^{3(1+w)}] = 0. \quad (14.15a)$$

This is simply integrated to give a relation for the evolution of the energy density

$$\rho a^{3(1+w)} = \text{const.}, \quad \rho(t) = \rho(t_0) \left(\frac{a(t_0)}{a(t)} \right)^{3(1+w)}. \quad (14.15b)$$

Here t_0 is some convenient time, such as the present. In general the energy density of a constituent decreases as the universe expands. In particular for matter the decrease is proportional to the inverse cube and for radiation it is proportional to the inverse fourth power of the scale factor. These are actually well-known classical properties of matter and radiation contained in an expanding volume, so this result of general relativity should be viewed as a verification of the consistency of the theory with classical physics.

The above result lets us write the energy density, matter plus radiation, as a simple sum; for the case of cold matter and radiation it is

$$\rho = \rho_m(t_0) \left(\frac{a(t_0)}{a(t)} \right)^3 + \rho_r(t_0) \left(\frac{a(t_0)}{a(t)} \right)^4, \quad (14.16a)$$

which we abbreviate and rewrite as

$$\rho = \rho_{m0} \left(\frac{a_0}{a} \right)^3 + \rho_{r0} \left(\frac{a_0}{a} \right)^4. \quad (14.16b)$$

Thus the density throughout time is simply related to the present densities and the scale factor; relations (14.16) are very elegant and useful. Recall also that the scale factor may be taken as unity at the present time, making (14.16) yet simpler looking.

Equations (14.16) are key relations in obtaining the master equation for the evolution of the scale factor that we will derive in the next section.

14.5 The Friedmann Master Equation

In this section we will obtain an equation that allows direct calculation of the scale factor in a form especially well-suited to some of the most physically interesting situations. We first rearrange the fundamental Einstein equation (14.5a) to give

$$a'^2 - \left(\frac{8\pi G}{3c^2} \right) \rho a^2 - \frac{\Lambda c^2}{3} a^2 + kc^2 = 0. \quad (14.17)$$

Then we substitute the density expression for matter and radiation from (14.16b) to get

$$a'^2 - \left(\frac{8\pi G}{3c^2} \right) \left[\rho_{m0} \left(\frac{a_0}{a} \right)^3 + \rho_{r0} \left(\frac{a_0}{a} \right)^4 \right] a^2 - \frac{\Lambda c^2}{3} a^2 + kc^2 = 0. \quad (14.18)$$

This is a rather simple first order differential equation for the scale factor. It can be put into more beautiful form by using the definition of the critical density, the vacuum density and the curvature density in (14.7) and (14.8). Using these we may put (14.18) into the form,

$$\begin{aligned} \frac{a'^2}{a^2} - \left[\Omega_{m0} \left(\frac{a_0}{a} \right)^3 + \Omega_{r0} \left(\frac{a_0}{a} \right)^4 + \Omega_{v0} + \Omega_{k0} \left(\frac{a_0}{a} \right)^2 \right] H_0^2 &= 0, \\ \Omega_{v0} \equiv \frac{\Lambda c^2}{3H_0^2}, \quad \Omega_{k0} \equiv -\frac{kc^2}{a_0^2 H_0^2}, \quad \Omega_{m0} \equiv \frac{8\pi G \rho_{m0}}{3c^2 H_0^2}, \quad \Omega_{r0} \equiv \frac{8\pi G \rho_{r0}}{3c^2 H_0^2}. \end{aligned} \quad (14.19)$$

This is quite useful and elegant: it is a first order differential equation for the scale factor in terms of powers of the scale factor. Moreover the various Ω coefficients denoted by a zero subscript can all be determined by observations of the present universe.

We will refer to (14.19) as the Friedmann master equation; however Friedmann's name has also been attached to various related equations, including (14.5).

There is one useful feature of the Friedmann master equation that is worth noting at this point. The various epochs in the evolution of the universe involve the scale factor going from very small to very large values. From (14.19) we see this means that over time, roughly speaking, the most important ingredients are radiation, then

cold matter, then curvature, then finally dark energy or the cosmological constant. This does not include the hypothetical epoch of inflation that we will discuss in a later chapter.

Appendix 1: The Einstein Tensor for the FLRW Metric

The Riemann and Ricci tensors were defined and discussed in Chaps. 8 and 10, and in particular the Ricci tensor is given in (8.27). The Einstein tensor which occurs on the geometric left side of the field equations was defined in terms of the Ricci tensor in (8.31). For the diagonal FLRW metric it is straight-forward to calculate the Ricci tensor and the Ricci scalar; the result for the nonzero components is (Schutz 2009)

$$\begin{aligned} R_{00} &= \frac{3a''}{ac^2}, & R_{11} &= \frac{-1}{1-kr^2} \left[2k + \frac{aa''}{c^2} + \frac{2a'^2}{c^2} \right], \\ R_{22} &= -r^2 \left[2k + \frac{aa''}{c^2} + \frac{2a'^2}{c^2} \right], & R_{33} &= -R_{22} \sin^2 \theta, \\ R &= R^\beta{}_\beta = \frac{6}{a^2} \left[k + \frac{aa''}{c^2} + \frac{a'^2}{c^2} \right]. \end{aligned} \quad (14.20)$$

From the Ricci tensor and Ricci scalar the 0,0 and 1,1 components of the Einstein tensor are

$$G_{00} = -\frac{3a'^2}{a^2c^2} + \frac{3k}{a^2}, \quad G_{11} = \frac{1}{1-kr^2} \left[k + \frac{2aa''}{c^2} + \frac{a'^2}{c^2} \right]. \quad (14.21)$$

Finally we raise an index and obtain the mixed index forms

$$G^0{}_0 = -3 \left[\frac{a'^2}{a^2c^2} + \frac{k}{a^2} \right], \quad G^1{}_1 = - \left[\frac{k}{a^2} + \frac{2aa''}{a^2c^2} + \frac{a'^2}{a^2c^2} \right]. \quad (14.22)$$

This verifies the Einstein tensor (14.4) in the text. Note that in the mixed index form there is no explicit spatial dependence in the Einstein tensor, which is a convenient feature. The other components of the field equations are either identically zero or the same as the above. See Exercise 14.2.

Exercises

- 14.1 Calculate the affine connections for the FLRW metric. Alternatively see Misner (1973) and Schutz (2009).
- 14.2 Verify the calculation of the Ricci and Einstein tensors for the FLRW metric in the Appendix or see Schutz (2009) and Misner (1973). Show that the other components of the field equations are either identically zero or redundant. In particular show $G^1{}_1 = G^2{}_2 = G^3{}_3$.

- 14.3 Consider a star orbiting near the edge of a galaxy in a circular orbit, with the mass of the galaxy concentrated in the central bulge. What is the relation between the orbital velocity and radius of the orbit?
- 14.5 Now suppose that the galaxy is dominated by dark matter distributed spherically symmetrically as in Fig. 14.1 with density $\rho(r)$. What is the relation between the orbital velocity and radius of the orbit? In the special case that the velocity is constant show that the density distribution is proportional to $1/r^2$. What is the total mass of the dark matter in the galaxy? Is this a problem?
- 14.6 Look up a reference on the work of Zwicky, then work out the way that the random motion in a cluster of galaxies (velocity dispersion) can determine their mass (Zwicky 1933; Wiki DM).

Chapter 15

Solutions for the Present Universe



Abstract The universe is presently dominated by vacuum or dark energy, described by the cosmological constant Λ , and cold dark matter; it is thus referred to as the Λ CDM universe. In the spatially flat case the Friedmann master equation may be solved for the two ingredients separately and also for both together; the combined solution is remarkably simple and useful in understanding properties of the universe.

15.1 The Positive Cosmological Constant

Preparatory to solving the master dynamical equation (14.19) we will show that the cosmological constant must be positive and establish an important relation among the cosmic fluid constituent densities.

As we have indicated in previous chapters, observations of distant supernovae show that the universe is accelerating, with a negative deceleration parameter of about $q_0 = -0.55$. This clearly implies that a'' is positive according to the definition (13.28). It is easy to show that the universe can only be accelerating if the cosmological constant is positive. To see this we differentiate the master equation (14.19) and find that the second derivative of the scale factor at the present time is

$$a'' = a_0 \left[\Omega_{V0} - \left(\frac{\Omega_{m0}}{2} + \Omega_{r0} \right) \right] H_0^2, \quad \Omega_{V0} \equiv \frac{\Lambda c^2}{3H_0^2}. \quad (15.1)$$

Since this second derivative is positive Ω_{V0} must also be positive and so must the cosmological constant Λ .

Equation (15.1) may be put into a simpler and useful form in terms of the deceleration parameter defined in (13.28). We find

$$q_0 = \left(\frac{\Omega_{m0}}{2} + \Omega_{r0} \right) - \Omega_{V0}. \quad (15.2)$$

This is an important relation between measurable quantities in the real world. It is consistent with the present values of about $q_0 = -0.55$, $\Omega_{V0} = 0.70$, $\Omega_{m0} = 0.30$, $\Omega_{r0} = 0$.

15.2 Complete Solution of the Friedmann Master Equation

In one sense the Friedmann equation (14.19) is immediately solvable, that is by quadratures. We need simply solve for the positive first derivative of the scale factor and integrate (14.19). We thereby obtain the solution according to

$$\frac{da}{dt} = a \left[\Omega_{m0} \left(\frac{a_0}{a} \right)^3 + \Omega_{r0} \left(\frac{a_0}{a} \right)^4 + \Omega_{V0} + \Omega_{k0} \left(\frac{a_0}{a} \right)^2 \right]^{1/2} H_0, \quad (15.3a)$$

$$\begin{aligned} & \int_0^a \frac{da}{a} \left[\Omega_{m0} \left(\frac{a_0}{a} \right)^3 + \Omega_{r0} \left(\frac{a_0}{a} \right)^4 + \Omega_{V0} + \Omega_{k0} \left(\frac{a_0}{a} \right)^2 \right]^{-1/2} \\ &= H_0 \int_0^t dt = H_0 t. \end{aligned} \quad (15.3b)$$

Here we have assumed that the scale factor is zero at time zero. While this is a complete solution it is not the most revealing form of solution. In the following sections we will obtain useful analytic forms of the solution for various epochs which are dominated by only one or two terms in the square bracket in (15.3).

Notice however that (15.3b) is in convenient form for numerical solution. One need only insert appropriate values of the present density ratios and let the computer integrate. Also note that there is a rather informative mechanical analog to the Friedmann master equation which we can use for a qualitative analysis. This is discussed in Appendix 1.

In the rest of this chapter we will usually take advantage of our freedom in choosing the value of the scale factor at some convenient time, and will take the value at present to be unity, $a_0 = a(t_0) = 1$. This simplifies the look of the equations.

15.3 Cosmological Constant Dominance

First we consider the present universe, which is largely dominated by the cosmological constant, or dark energy. For this we use the Friedmann master equation in the form (14.18), which is equivalent to (14.19). Neglecting the matter and radiation terms.

$$a'^2 - (\Lambda c^2/3)a^2 = -kc^2. \quad (15.4)$$

This is simple enough that we may solve by inspection.

For zero curvature the solution of (15.4) is an exponential, which we write with an arbitrary constant t_e as

$$a = a(t_e)e^{\sqrt{\Lambda/3} c(t-t_e)}, \quad k = 0. \quad (15.5)$$

Thus the scale factor is that of de Sitter space, which we discussed in Sect. 13.4, and the Hubble constant is $H_0 = \sqrt{\Lambda/3} c$. For positive curvature the solution is a hyperbolic cosine, which we write with an arbitrary constant t_+ as

$$a = \sqrt{3k/\Lambda} \cosh\left[\sqrt{\Lambda/3} c(t - t_+)\right], \quad k > 0. \quad (15.6a)$$

For negative curvature the solution is a hyperbolic sine, which we write with an arbitrary constant t_- as

$$a = \sqrt{3|k|/\Lambda} \sinh\left[\sqrt{\Lambda/3} c(t - t_-)\right], \quad k < 0. \quad (15.6b)$$

Notice that for asymptotically large times all three are exponential functions.

The three times t_e , t_+ , t_- are arbitrary constants of integration. They could all be chosen to make the scale factor equal to unity at the present time, as we usually do. However to display the solutions for this one case we will take a different approach. We choose the constant t_e to be the present time t_0 but choose t_+ and t_- so that all three solutions are asymptotically equal for very large times. For the positive curvature case the necessary choice of t_+ is determined by setting the large time behavior equal to the exponential (15.5), leading to

$$e^{\sqrt{\Lambda/3} c(t_0 - t_+)} = 2a_0\sqrt{\Lambda/3k}, \quad c(t_0 - t_+) = \sqrt{3/\Lambda} \log(2\sqrt{\Lambda/3k}a_0). \quad (15.7)$$

The reader may work out the analogous relation for negative curvature. This choice of normalization is reasonable since these models are appropriate to the universe at late times when the scale factor is large. Indeed the curvature k is likely to be unimportant for the real universe at the present time.

Figure 15.1 shows how the three cases we have discussed, exponential and hyperbolic sine and cosine, become asymptotically equal at large times. It is amusing that the scale factor has such a simple form for large times, that of de Sitter space. See also Sect. 16.7 for a different time coordinate for the case of $k = 0$.

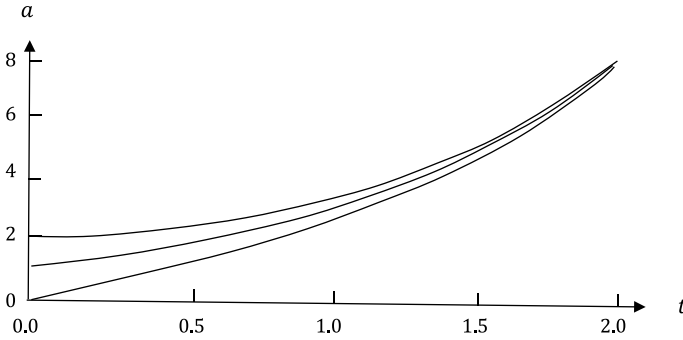


Fig. 15.1 The three solutions for a universe with only a cosmological constant are the exponential and sinh and cosh curves. With appropriate normalization they are asymptotically equal for large times. The units of the scale factor are $\sqrt{3|k|/\Lambda}$ and the units of time are $\sqrt{\Lambda/3}c$

15.4 Matter Dominance

By matter dominance we mean the epoch in which pressure can be neglected, and in which the cosmological constant is less important than matter density. For the real world this corresponds roughly to times from near the beginning of the universe, a few hundred thousand years, until about 7 billion years, as we will discuss below. This was the first case studied by Friedmann, long before the observational discovery that the cosmological constant is positive and the universe is accelerating (Adler 1975).

For the matter epoch the solution in the integral form (15.3b) is most useful,

$$\int_0^a da \left[\frac{\Omega_{m0}}{a} + \Omega_{k0} \right]^{-1/2} = H_0 t, \quad \Omega_{k0} \equiv -\left(\frac{c^2}{H_0^2 a^2} \right) k. \quad (15.8a)$$

As usual we have taken $a_0 = 1$. For the case of no curvature, $k = 0$, the integration is immediate and simple

$$\int_0^a \frac{\sqrt{a}}{\sqrt{\Omega_{m0}}} da = \frac{2}{3} \frac{a^{3/2}}{\sqrt{\Omega_{m0}}} = H_0 t. \quad (15.8b)$$

This gives immediately the scale factor and the age of the universe,

$$a = \left(\frac{t}{t_0} \right)^{2/3}, \quad t_0 = 1/\sqrt{\Omega_{m0}} H_0, \quad \text{for } k = 0. \quad (15.9)$$

Notice that this solution is explosive for early times in that the derivative of the scale factor is infinite at zero. Notice also that it is obvious that for early times and small

scale factor the integral in (15.8a) is dominated by the matter term, so the solutions for all k will behave like (15.9).

For nonzero values of k the integral in (15.8a) may also be evaluated easily. For $k > 0$ and $\Omega_{k0} < 0$, the 3-sphere, the curvature density is negative and the we have

$$\int_0^a da \left[\frac{\Omega_{m0}}{a} - |\Omega_{k0}| \right]^{-1/2} = H_0 t \quad (15.10)$$

From integral tables we obtain

$$\frac{1}{\sqrt{|\Omega_{k0}|}} \left[D \sin^{-1}(\sqrt{a/D}) - \sqrt{a(D-a)} \right] = H_0 t, \quad (15.11)$$

$$D = \left| \frac{\Omega_{m0}}{\Omega_{k0}} \right|, \quad \Omega_{k0} < 0.$$

From this somewhat cumbersome expression we may plot the behavior of the scale factor and see that it is qualitatively as shown in Fig. 15.2. The curve is known as a *cycloid* and is explored further in Exercise 15.1. In this model the scale factor increases to a maximum value D and then decreases to zero after a finite time; the universe does *not* expand forever.

For $k < 0$ and $\Omega_{k0} > 0$, the 3-pseudosphere, the same manipulations give

$$\frac{1}{\sqrt{\Omega_{k0}}} \left[\sqrt{a(D+a)} - D \sinh^{-1}(\sqrt{a/D}) \right] = H_0 t, \quad (15.12)$$

$$D = \frac{\Omega_{m0}}{\Omega_{k0}}, \quad \Omega_{k0} > 0.$$

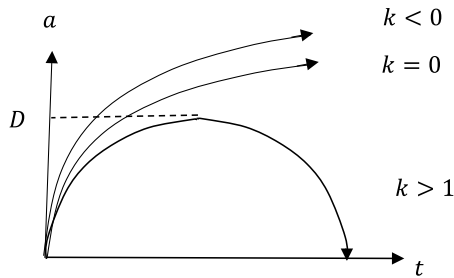


Fig. 15.2 Qualitative behavior of the scale factor for negative, zero, and positive curvature parameter k . See Appendix 1 for comments on a mechanical analogy

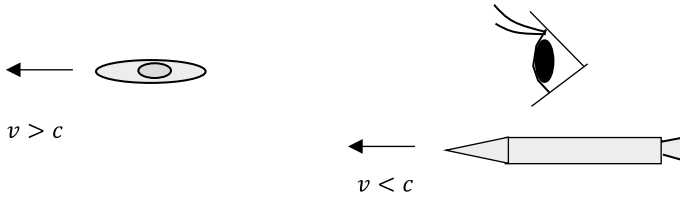


Fig. 15.3 The distant galaxy may recede at velocity greater than c , but the rocket may not pass by us at greater than c . The galaxy will not be visible

As with the 3-sphere we may plot the behavior of the scale factor and see that it behaves as shown in Fig. 15.2. We may call the curve a *pseudocycloid*. It is explored further in Exercise 15.2. In this model the scale factor increases forever.

The three solutions (15.9) and (15.11) and (15.12) are the classic Friedmann solutions. They were the first realistic cosmological solutions and indeed were the favored solutions before the discovery of the accelerating universe and the positive cosmological constant.

Example 15.1 Faster Than Light? Consider two galaxies separated by a constant co-moving coordinate distance σ and physical distance $a(t)\sigma$. The velocity of separation in the flat matter dominated universe is

$$v = a'\sigma = \left(\frac{2}{3}\right) \frac{\sigma}{t_0^{2/3} t^{1/3}}. \quad (15.13)$$

For early times (and small a) this is greater than c , and even becomes infinite! This may be somewhat disturbing, but is not really a violation of any principle of relativity. In all of special relativity, general relativity and cosmology the physical velocity of light is the invariant c , and no two objects may pass each other at a velocity greater than c . Matter in a distant galaxy is not included in this dictum! See Fig. 15.3.

Indeed the observable result of the rapid expansion of the universe is that a galaxy moving away from us at greater than c simply cannot be seen. We will return to this question when we study horizons in Chap. 16, but at this point the reader should convince himself no conceptual inconsistency results from such motion.

15.5 The LCDM Universe

Now we will combine the material of the preceding two sections and study a model universe dominated by the cosmological constant Λ and cold matter; the cold matter is mainly cold dark matter, CDM; it is variously called the Λ CDM model, or the LCDM model, or the standard model. Because it appears to be consistent with all observations it is also widely called the concordance model.

For this model our task is to evaluate the integral in (15.3b) without the radiation term, that is

$$\int_0^a da \left[\Omega_{V0} a^2 + \frac{\Omega_{m0}}{a} + \Omega_{k0} \right]^{-1/2} = H_0 t. \quad (15.14)$$

This is an implicit solution, but this form of solution is not particularly useful since the integral does not involve elementary functions; it can of course be solved numerically. But for the case of zero curvature the integration can be done in terms of elementary functions. It now appears that curvature of the real universe is either zero or quite small so we will focus on this favored case, and since the result is important we will do the integration in explicit detail.

To do the integral in (15.14) we first simplify it by introducing a new dimensionless time τ and scale factor y . The substitutions and the resulting equation are

$$\tau = \sqrt{\Omega_{V0}} H_0 t = \sqrt{\Lambda/3} ct, \quad a = \left(\frac{\Omega_{m0}}{\Omega_{V0}} \right)^{1/3} y, \quad \tau = \int_0^y \frac{dy}{\sqrt{y^2 + 1/y}}. \quad (15.15)$$

Then we make another substitution for the variable of integration $y^3 = x^2$ and obtain

$$\tau = \frac{2}{3} \int_0^x \frac{dx}{\sqrt{1+x^2}} = \frac{2}{3} \log(x + \sqrt{1+x^2}), \quad x^2 = y^3. \quad (15.16)$$

Fortunately this may be easily inverted to give $x(\tau)$ and thus $y(\tau)$ from (15.15),

$$\sqrt{1+x^2} = e^{3\tau/2} - x, \quad x(\tau) = \sinh(3\tau/2), \quad y(\tau) = \left[\sinh\left(\frac{3}{2}\tau\right) \right]^{2/3}. \quad (15.17)$$

Finally, in terms of the original functions and parameters, the scale function is

$$a(t) = \left(\frac{\Omega_{m0}}{\Omega_{V0}} \right)^{1/3} \left[\sinh\left(\frac{\sqrt{3\Lambda}}{2} ct\right) \right]^{2/3}. \quad (15.18)$$

Figure 15.4 shows the behavior of the scale factor as well as the asymptotic forms for small and large times.

Note that setting the scale factor equal to 1 at the present time gives the age of the universe for the LCDM model; we will return to this in Chap. 16. See also Exercise 15.7.

Equation (15.18) is a remarkable result. It is the exact solution of the dynamical equations for the currently favored model of the real universe, the flat LCDM model.

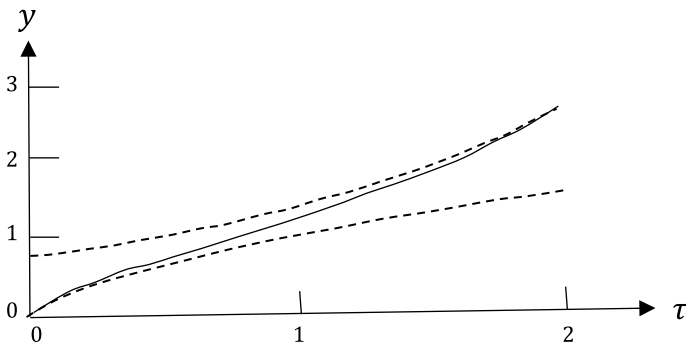


Fig. 15.4 The solid curve is the LCDM scale factor in (15.18) and the dashed curves are the large and small time limits. The axes are labelled with the scaled variables in (15.15).

It is believed to describe the universe for most of its history, from a few hundred thousand years after its beginning to the present day, and into the indefinite future. For earlier times we must consider radiation and hot matter as important ingredients of the universe, which we will do in later chapters. In Chap. 16 we will further discuss some interesting properties of the flat LCDM universe based largely on (15.18).

If the curvature of the universe is not exactly zero the integral in (15.14) does not reduce to an elementary function but it can be evaluated approximately with Ω_{k0} taken as small. As we should expect there is no way to determine observationally if Ω_{k0} is exactly zero; thus we can only say that we live in a nearly flat universe (Adler 2005). See Exercise 15.8.

Appendix 1: A Mechanical Analogy

Equation (14.19) may be analyzed qualitatively using a mechanical analogy. Indeed the analysis is quite general and nicely illustrates the behavior of the universe with time. We first rearrange (14.19) slightly and compare it with the equation describing a projectile of unit mass $m = 1$ moving radially in a potential $V(r)$,

$$\frac{a'^2}{2} - \frac{1}{2} \left[(\Omega_{m0} H_0^2) \frac{1}{a} + \frac{\Lambda c^2}{3} a^2 \right] = -\frac{kc^2}{2} \Leftrightarrow \frac{mr'^2}{2} + V(r) = E. \quad (15.19)$$

As usual we have taken the present scale factor to be unity and neglected the radiation density, which is small for the present universe. The equations are the same if the mechanical analog quantities are related by

$$a \Leftrightarrow r, \quad V(r) \Leftrightarrow -\frac{1}{2} \left[(\Omega_{m0} H_0^2) \frac{1}{a} + \frac{\Lambda c^2}{3} a^2 \right], \quad E \Leftrightarrow -\frac{1}{2} kc^2. \quad (15.20)$$

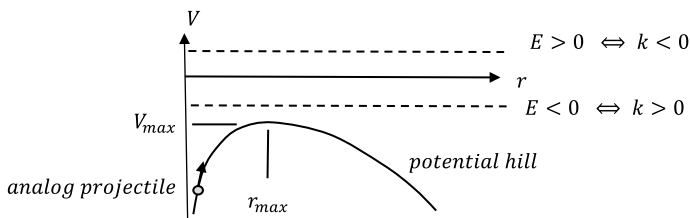


Fig. 15.5 Qualitative sketch of the effective potential for the mechanical analogy

That is, the universe expands like a projectile moving radially in a potential composed of two terms: one term is an attractive Newtonian potential and the other is a repulsive quadratic potential—that is a harmonic oscillator potential with the wrong sign. This potential is shown in Fig. 15.5.

We suppose the projectile starts at small r with a positive velocity and total energy E as shown in the figure. The position and maximum of the potential are, from (15.20),

$$r_{\max} = \left(\frac{3\Omega_{m0}H_0^2}{2\Lambda c^2} \right)^{1/3}, \quad V_{\max} = -\left(\frac{9}{32} \right)^{1/3} (\Lambda H_0^4 \Omega_{m0}^2 c^2)^{1/3}. \quad (15.21)$$

Consider a projectile having negative energy, corresponding to $k > 0$. From the figure it is clear that it will move upward from its beginning position to some maximum and fall back if the total energy is less than the maximum of the potential $V_{\max}$. If this criterion for recontracting is satisfied the universe expands to a maximum size, and falls back to zero for a “big crunch” qualitatively similar to the $k = 1$ universe of Sect. 15.4. If the observations discussed in Sect. 15.1 indicating an accelerating universe are correct then this case is in fact ruled out.

For the critical value of $E = V_{\max}$ the universe has interesting behavior: it expands to its maximum size and stays there forever. But this situation is clearly unstable as is apparent from the mechanical analog and Fig. 15.5, so in fact we expect it to eventually contract or expand further. This static solution was the first cosmology proposed by Einstein, but is now of only historical interest due to its instability. See Exercise 15.5.

In cases other than the above two, the universe begins with small size and expands to $a \rightarrow \infty$, which is apparently what nature has chosen. For late times and large a the behavior is exponential as discussed in Sect. 15.3.

Appendix 2: Newtonian View of Dark Energy

Dark energy arises naturally in the context of general relativity theory. It is associated with the cosmological constant as we have discussed in detail, and it has the important feature of being constant in both space and time. Due to dark energy the universe

undergoes accelerated expansion, so there is clearly a repulsive force associated with it. The repulsive force is perhaps best seen from the Kottler metric, which is also called the Schwarzschild—de Sitter metric (Adler 1975). The Kottler metric describes the field of a spherical mass distribution in a universe containing dark energy; it can be derived much as we derived the Schwarzschild metric in Chap. 9 and is

$$ds^2 = g_{00}c^2dt^2 - g_{00}^{-1}dr^2 - r^2(d\theta^2 + \sin^2\theta d\varphi^2), \quad g_{00} = 1 - \frac{2m}{r} - \frac{r^2}{R_d^2},$$

$$2m \equiv \frac{2GM}{c^2} \text{ Schwarzschild radius}, \quad R_d^2 \equiv \frac{3}{\Lambda} \text{ de Sitter radius.} \quad (15.22)$$

For small distances it approaches the Schwarzschild metric as we should expect.

Recall from Chap. 7 that general relativity reduces to Newtonian gravitational theory in the low velocity and weak field limit, or classical limit, if (7.23) is valid, which we repeat here

$$g_{00} = 1 + \frac{2\phi}{c^2}. \quad (15.23)$$

Comparing (15.23) with (15.22) we see that there are two corresponding classical potentials and forces produced by the mass and cosmological constant, which we can call the Newtonian and dark energy potentials and forces; they are

$$\frac{\phi_N}{c^2} = -\frac{m}{r}, \quad \frac{\phi_{DE}}{c^2} = -\frac{r^2}{2R_d^2}, \quad \text{classical potentials,} \quad (15.24a)$$

$$F_N = -\frac{m}{r^2}c^2, \quad F_{DE} = \frac{r}{R_d^2}c^2, \quad \text{classical forces per unit mass.} \quad (15.24b)$$

The ratio of the forces at a given r is a convenient dimensionless measure of their relative importance,

$$\varepsilon = \frac{r^3}{mR_d^2} = \left(\frac{r}{m}\right)\left(\frac{r}{R_d}\right)^2. \quad (15.25)$$

Thus a test particle at $r^3 = mR_d^2$ will feel no radial force, corresponding to the maximum radius for circular orbits.

As an example application consider a system of test particles in the potentials (15.24a). This might represent an approximate model for a cluster of galaxies. The Virial Theorem of classical mechanics can be applied to show that the root mean square velocity is given by

$$\langle v^2 \rangle = \left(\left\langle \frac{m}{r} \right\rangle - \left\langle \frac{r^2}{R_d^2} \right\rangle \right) c^2, \quad (15.26)$$

where the brackets indicate an average (Goldstein 1980). The last equation implies that the relative importance of the dark energy repulsive force to the attractive Newtonian force may be estimated as

$$\left\langle \frac{r^2}{R_d^2} \right\rangle / \left\langle \frac{m}{r} \right\rangle = \left(\frac{r_{\text{ch}}}{m} \right) \left(\frac{r_{\text{ch}}}{R_d} \right)^2 = \varepsilon_{\text{ch}}, \quad r_{\text{ch}}^3 = \langle r^2 \rangle \left\langle \frac{1}{r} \right\rangle. \quad (15.27)$$

Note that, with r_{ch} as defined in the above, the ratio ε_{ch} has the same form as ε in (15.25). This ratio may be evaluated for galactic clusters, the largest bound structures in the universe, to determine the relative effects of dark energy, which are small. For an example see Exercise 15.9.

It is interesting to see how dark energy might fit ab initio into classical gravitational theory as discussed in Chap. 7. As we emphasized above, dark energy is characterized by a constant density, so we consider Poisson's equation with a constant source

$$\nabla^2 \phi = s, \quad s = \text{constant source} \quad (15.28)$$

The spherically symmetric solution to this is

$$\phi = \frac{s}{6} r^2. \quad (15.29)$$

Thus we see that Poisson's equation produces the same potential as general relativity in the classical limit if we take the source s to be

$$s = -\frac{3c^2}{R_d^2} = -\Lambda c^2 = -8\pi G(\rho_{\text{DE}}/c^2), \quad \rho_{\text{DE}} = \text{dark energy density}. \quad (15.30)$$

That is, the source is negative $4\pi G$ times twice the dark energy mass density. In the context of classical physics it is hard to justify such a negative source, whereas in general relativity theory it arises naturally. See Exercise 15.10 as to how the factor of -2 explicitly arises. The negative sign is a peculiar and important feature of dark energy because of its effect on the accelerated expansion of the universe.

Finally we observe that the sort of correspondence between general relativity and classical gravity theory we have discussed in Chap. 7 and in this appendix allows a pedagogical development of cosmological theory based largely on classical physics (Liddle 2003). However one needs to postulate the sign of the repulsive force due to dark energy.

Appendix 3: Some Discarded Cosmological Models

In this chapter we have discussed models that are currently considered viable. We will mention only briefly some well-known models that are no longer considered viable.

Static models were first studied, by Einstein and others, before the expansion of the universe was discovered by Hubble and other astronomers. They are therefore no longer of physical interest.

Some models are not homogeneous and isotropic. One notable example is that of Godel, which has a preferred axis and rotates. It is of philosophical and theoretical interest since one must ask “With respect to what can the entire universe rotate?” (Godel 1949; Adler 1975). The Godel model also has interesting and peculiar causality properties (Hawking 1973). However the model has the fatal flaw of having no Hubble expansion and is not considered a viable model of the actual universe.

A steady state model was popular some decades ago, in which the universe expanded but did not change over time. Spontaneous creation of matter was necessary in this model, and it had no big bang by intent. The discovery of the cosmic microwave background radiation left over from the big bang greatly reduced interest in this model and it is no longer in the mainstream of cosmology (Bondi 1948; Hoyle 1948; Liddle 2003).

The de Sitter model has an exponential expansion and no big bang. It is the asymptotic limit of the Λ CDM model in the distant future when the scale factor becomes very large, as we discussed in the preceding sections. We will return to it as a mathematical guide when we discuss inflation in the very early universe; it is widely believed that during inflation the universe was dominated by a field or fluid that behaved much like extremely dense dark energy. As a complete model of the present universe however it is no longer viable.

Exercises

- 15.1 Plot the cycloid in (15.11) and show that it looks like the curve in Fig. 15.2. Look up the name cycloid to see how it relates to the motion of a point on the edge of a wheel.
- 15.2 Plot the pseudo-cycloid in (15.12) and show that it looks like the curve in Fig. 15.2.
- 15.3 If a galaxy has a redshift of z in a flat matter dominated universe how far away is it?
- 15.4 Work out properties of the static model from (15.3), by setting the scale factor equal to a constant. What values of the curvature parameter k are allowed? How does the scale factor depend on the density parameters?
- 15.5 Show explicitly that the static model is not stable. Appendix 1 should be of help.
- 15.6 Solve (15.3a) numerically for the Λ CDM model with a small nonzero k .
- 15.7 Equation (15.18) can be used to determine the age of the universe in the Λ CDM model. Do this by setting the scale factor equal to 1 in (15.18) and

thereby calculate the time t_0 as

$$t_0 = \frac{2}{3H_\infty} \text{Arccsinh} \sqrt{\frac{\Omega_{v0}}{\Omega_{m0}}}, \quad H_\infty = \sqrt{\Lambda c^2/3}.$$

See Sect. 16.3 for an equivalent way to calculate the age.

- 15.8 In (15.14) take the curvature parameter Ω_{k0} to be small and evaluate the integral approximately (Adler 2005). This is the case of the nearly flat universe; it may be compared with observations in order to place limits on the curvature parameter k .
- 15.9 As an example of the Newtonian approach to dark energy in Appendix 2 consider the Coma cluster of galaxies; take its characteristic size and mass and the de Sitter radius very roughly to be

$$r_{\text{ch}} \cong 0.95 \times 10^{20} \text{ km}, \quad m \cong 1.0 \times 10^{15} \text{ km}, \quad R_d \cong 1.6 \times 10^{23} \text{ km}.$$

Using (15.27) show that the effect of dark energy on the rms velocity is only a few percent. Note that r_{ch}/m is about 10^5 , which is observed to be approximately the same for all galaxies and clusters in the universe!

- 15.10 We have seen in Chap. 7 that Newtonian gravity occurs in the limit of general relativity when the source is slowly moving mass with density ρ and zero pressure. Include pressure in the calculation and show that the appropriate source in the classical limit is $\rho + 3p$; that is, pressure is a source of gravity. Hence for dark energy with $p = -\rho$ the correct source density is $-2\rho_{\text{DE}}$ exactly as we obtained in (15.30).

Chapter 16

Some Properties of the LCDM Universe



Abstract In the last half century or so there have been great advances in cosmological observations, making cosmology a thriving mix of theory and observations. In this chapter we will discuss how we understand from LCDM theory the expansion of the universe, its age, and the behavior of its dominant constituents, the dark energy and cold dark matter. We will also relate and compare the theory with the observed values of various cosmological parameters such as the Hubble constant. Despite the impressive agreement between theory and observation the basic physical nature of the dark matter is not yet known.

16.1 Diverse Cosmological Observations

In its first half century relativistic cosmology was based on very few observations, some of which we discussed in Chaps. 13 and 14. Sandage is quoted as saying that observational cosmology was the search for two numbers, the Hubble constant and the deceleration parameter (Sandage 1961). That viewpoint changed greatly in the latter half of the twentieth century and the early twenty-first century. The field of observational cosmology exploded and is now in the mainstream of physics and astronomy; there are many ways being used to gather information about the large-scale universe, many researchers are doing observations and theory and simulations, and there are many references to their work. This book is about the ideas and basic mathematics of the theory so we can only give a sketch of the observations being done and the kind of information they give us about the universe. Fortunately there are numerous references for the interested reader on the observations; a rather extensive set of references is given in the NASA website (NASA 2019).

Here are just some examples of interesting ways to gather information about the large-scale universe:

1. *Velocity and distance measurements of galaxies and supernovas*: Since Hubble's work early in the twentieth century a great deal of effort has gone into extending the distance scale to obtain more accurate values for the Hubble constant, the most important number in cosmology, and also the deceleration parameter that we discussed in Chap. 13.

We have already discussed the measured values of the Hubble constant in Appendix 1 in Chap. 13. In short summary the primary measurements involve a “distance ladder,” in which the distance to relatively nearby objects, in particular Cepheid variable stars, is measured using parallax; the Cepheids then provide a standard candle for larger distances since their periodicity is related to their brightness in a known way; finally the Cepheids in distant galaxies allow a determination of the distance to supernovas and the supernovas in turn serve as standard candles for greater distances (Weinberg 1972; Ohanian 1994).

The objects first studied by Hubble and other early workers were galaxies, but supernovas and red giant stars have allowed much larger distances to be measured. This has allowed researches to determine the age of the universe and to discover the acceleration of the universe that implies the existence of dark energy. We will say more about these concepts later in this chapter.

2. *Velocity measurements of galaxies in clusters and of stars in outer portions of galaxies:* We discussed this type of measurement briefly in Sect. 14.3. The earliest evidence for dark matter was in the random velocities of galaxies in galactic clusters. Such velocities are directly related to the gravitational potential in their region; already in the 1930s Zwicky realized that there must be much more matter in many such clusters than was indicated from the luminous matter (Zwicky 1933). Several decades later Rubin used measurements of the velocities of stars in the outer regions of individual galaxies to reach a similar conclusion (Rubin 1995). Since galactic clusters are so much larger than galaxies such measurements on both scales are very important; they are probably the simplest evidence for the existence and universality of dark matter.
3. *Gravitational lensing by dark matter:* In Chap. 10 we discussed the bending of light in a gravitational field. In fact density distributions of matter can be analyzed by detecting the light that passes by them or through them, much as the shape of a magnifying glass can be determined by the way it focuses light. Einstein was the first to point out the existence of such lensing; he thought in terms of lensing by stellar size objects, whereas gravitational lensing as currently used makes use of light lensed by galaxy size distributions of matter, and in particular the dark matter that we mentioned in Chap. 14 (Schneider 1992).

Another application of gravitational lensing is in measuring cosmological distances and parameters, called cosmography. Consider a source galaxy whose light passes by and is deflected by a galaxy that is closer to us. In general we will see several images of the source galaxy, and the angles between the images will give us information on the distances involved (Narayan 1997; Schneider 1992). If the light from the source galaxy varies we will see those variations at different times. The combination of the angles between the images and the time delay of the variations can give us information on the distance to the source and cosmological parameters, in particular the Hubble constant (Chen 2019).

4. *Spectrum of the cosmic microwave background:* We have already mentioned the CMB in Chap. 13; it is interpreted as the dim afterglow of the radiation fireball of the early universe. In the next chapter we will discuss how the temperature and spectrum of the fireball and its present afterglow are determined; the spectrum is

very nearly that of an isothermal black body, but with small and very important variations. The spectrum was first measured by the Cosmic Background Explorer (COBE) satellite and found to fit the black body spectrum extremely well, to about a part in 10^5 (Boggess 1992). That alone was very strong evidence for the big bang paradigm. Since then the small variations in the spectrum have been accurately measured by the Wilkinson Microwave Anisotropy Probe (WMAP) and Planck satellites and used as tests of theories of the present and early universe—and also the very early universe (Bennet 2003; Planck 2018).

The detailed shape of the CMB spectrum depends on a small number of cosmological parameters, such as the Hubble constant, as well as the constituents and dynamics of the universe at about the time of emission. In particular the position and spacing of some peaks in the spectrum can be understood in terms of density oscillations or standing “acoustic” waves at the time of emission. Combined with a theoretical calculation of the speed of sound in the material this translates to information on the wavelength of such oscillations. Putting this information together we can determine the values of the parameters by fitting the theoretical CMB spectrum to the observed spectrum. (The fitting of the spectrum can be done using openly available programs such as CMBFAST.) One important result is that the Hubble constant may be accurately determined from the CMB spectrum as we discussed in Appendix 1 in Chap. 13 (Planck 2018). See also Sect. 17.4.

The dominant theories of the very early universe involve a very large and rapid expansion of the universe called *inflation*, which ended when the radiation era began. Processes during inflation also affect the CMB spectrum in interesting ways. We will discuss the motivation for such theories in Chaps. 17 and 18, and give a rough sketch of how they work in Chap. 19.

5. *Gravitational waves from black hole and neutron star mergers:* We discussed in Chap. 11 the gravitational waves from binary black holes and neutron stars spiraling in to merge. The inverse distance dependence of the amplitudes of such waves can be used to estimate their distance from us. The frequency and the change in the frequency over the course of the merger also provide additional information as we discussed in Sects. 11.5 and 11.6. As the binaries radiate energy the orbit decays and the frequency increases in a predictable way, that is a chirp, that depends on the binary masses, so the systems thus act as “standard sirens,” analogous to the standard candles used for traditional distance measurements. The amplitude and the dependence of frequency on time allow us to measure the Hubble constant independently of any other measurements (Schutz 1986; Holz 2018).

Such measurements of the Hubble constant using the binary neutron star merger GW170817 give a value that is consistent with those obtained in other ways, in particular those using the distance ladder method discussed above (Abbott 2019; Abbott 2017; Fischbach 2018). This was discussed in Appendix 1 in Chap. 13. The accuracy of such measurements at present is not comparable to distance ladder and CMB methods, but is expected to increase as more events are detected.

6. *Large scale structures*: The spectrum of the CMB is one example of large-scale structure in the early universe. The distribution of matter at later times is another, and depends on the contents and cosmological parameters of the universe. Intuition might lead us to expect that a random but roughly elliptical blob of matter would first collapse toward a line, and then that line would collapse toward a point. This is generally born out in the observations and simulations, but the detailed large-scale structure is a very different matter. The results of computer simulations can be found on the internet, and show a rich tapestry of filaments and blobs and voids, which are observable in the real universe (Cosmicweb 2019). Note that there is considerable overlap between the ideas of large-scale structure and gravitational lensing since the dark matter that produces the lensing constitutes much of the mass of the universe.
7. *Some possible further observations*: In addition to the types of current observations above we discuss in the remainder of this section some potential observations of future interest.

The light elements in the universe today were formed in the first few minutes after the big bang; the heavier elements were formed later in the interiors of stars, or in the collisions of neutron stars. The formation of the light elements is well understood theoretically and accurately predicted as a function of properties such as the ambient temperature and nucleon abundances (Weinberg 1988; WMAP 2010). Thus observations of the present element abundances serve as a test of conditions in the early universe. We will say a bit more about this big bang nucleosynthesis (BBN) in Chap. 18.

The polarization pattern of the CMB is quite interesting, in addition to its spectrum as noted above. Specifically, the pattern depends on processes that happened during inflation and can thus give information about gravitational waves produced during inflation and the inflationary energy scale (see Chap. 19). The detection and analysis of the polarization patterns is quite difficult, specifically the so-called B modes, but should be of great interest (cfa.harvard.edu 2019).

The universe is filled with extragalactic background light emitted by stars during the lifetime of the universe. High energy photons interact with this light and it thus attenuates gamma rays in their passage through space. The amount of attenuation depends on the expansion rate of the universe and the matter content along the line of travel of the gamma rays. As a result gamma ray telescopes can yield a measurement of the Hubble constant and the present matter density of the universe (Dominguez 2019).

The rate of change of the redshift is an interesting quantity in both theory and observations. The redshift z of receding galaxies is a fundamental property of the big bang cosmology and is measured very accurately. It is defined in terms of the scale factor in (13.22). The Friedmann equation (14.19) then determines the scale factor in the current universe as a function of time, depending on the Hubble constant and the present matter and vacuum energy densities. It is straightforward to calculate the time derivative of z from (14.19), and that yields a surprisingly simple expression, susceptible to testing. See Exercise 16.1. Given the present and expected accuracies

in measuring the redshift the time rate dz/dt might be measurable in the near future and thus yield information on the Hubble constant and the present matter density (Martins 2016; Eikenberry 2019).

There are many pulsars whose timing is being monitored to extreme accuracy. If a long wavelength gravitational wave passes by a number of such pulsars it will move them slightly and alter the timing of the electromagnetic pulses we receive from them; a system of pulsars can thus be used as a gravitational wave detector. This could be especially interesting for wavelengths much longer than can be detected using earth-based detectors such as LIGO or even LISA. Such measurements may become feasible in the near future (Lommen 2017).

16.2 Cosmological Parameter Values

In Table 16.1 we list approximate values for some important cosmological parameters. They are obtained from analyses of a variety of observation, some of which are discussed in the previous section. Some of the values are fairly rough estimates. Some are continually being remeasured, and some remain controversial—such as the Hubble constant. (It may be helpful to remember that $1 \text{ Mpc} = 3.09 \times 10^{22} \text{ m}$.)

Table 16.1 Values of diverse cosmological parameters and numbers

Hubble constant (our own error estimate)	$H_0 = 70 \pm 5 \text{ (km/s)/Mpc}$
Hubble time	$T_H = 1/H_0 = 1.40 \times 10^{10} \text{ year}$
Cosmological constant	$\Lambda = (0.964 \times 10^{10} \text{ ly})^{-2}$
Asymptotic Hubble function	$H_\infty = \sqrt{\Lambda c^2/3} = 59 \text{ (km/s)/Mpc}$
Asymptotic Hubble time, de Sitter time	$T_{dS} = 1/H_\infty = 1.67 \times 10^{10} \text{ year}$
Age of the universe	$t_0 = 1.35 \times 10^{10} \text{ year}$
Deceleration parameter	$q_0 = -0.55$
Cold dark matter energy density ratio, present	$\Omega_{dmo} = 0.25$
Baryonic matter energy density ratio, present	$\Omega_{b0} = 0.05$
Total matter energy density ratio, present	$\Omega_{m0} = 0.30$
Cosmological constant energy density ratio, present	$\Omega_{v0} = 0.70$
Radiation energy density ratio, at present	$\Omega_{r0} = 3.8 \times 10^{-5}$
Curvature effective energy density ratio, present	$\Omega_{k0} = 0$
Total energy density ratio according to GR	$\Omega_T = \Omega_m + \Omega_v + \Omega_k = 1.0$
Critical energy density	$\rho_{cr} = 8.28 \times 10^{-10} \text{ J/m}^3$
Time of matter and dark energy density equality	$t_e = 9.9 \times 10^9 \text{ year}$
Time of decoupling	$t_d = 380,000 \text{ year}$

Notice that we do not give error estimates for most of the parameters: this is because they are continually being re-evaluated. The reader who is interested in precise values should consult internet references for up-to-date values with error estimates; many such references can be found in the NASA website (NASA 2019).

16.3 The Hubble Function and the Age of the Universe

In the previous chapter we obtained the scale factor in (15.18) for the flat LCDM universe, the standard model of cosmology. It is a remarkable result in that it is believed to describe the actual universe from about the time of decoupling to the present and into the distant future. We repeat it here in the form

$$\begin{aligned} a(t) &= \left(\frac{\Omega_{m0}}{\Omega_{v0}} \right)^{1/3} \left[\sinh \left(\frac{3}{2} \sqrt{\frac{\Lambda c^2}{3}} t \right) \right]^{2/3} \\ &= \left(\frac{\Omega_{m0}}{\Omega_{v0}} \right)^{1/3} \left[\sinh \left(\frac{3}{2} H_\infty t \right) \right]^{2/3}. \end{aligned} \quad (16.1)$$

The constant $H_\infty = \sqrt{\Lambda c^2/3}$ is the value of the Hubble function in the asymptotically distant future; it is also called the inverse de Sitter time. From the scale factor (16.1) the Hubble function follows as

$$H(t) = \frac{a'}{a} = H_\infty \frac{\cosh(\frac{3}{2} H_\infty t)}{\sinh(\frac{3}{2} H_\infty t)} = H_\infty \coth \left(\frac{3}{2} H_\infty t \right). \quad (16.2)$$

For early times and late times $H(t)$ is approximately

$$H(t) = \frac{2}{3t} \text{ early times, } H(t) = H_\infty \text{ late times.} \quad (16.3)$$

For many years before the discovery of the accelerating universe, the universe was thought to have the early time Hubble function in (16.3), so the age of the universe was taken to be about 2/3 of the Hubble time or about 9.3 billion years. According to (16.2) we can calculate the age in the flat LCDM model by setting $H(t)$ equal to H_0 and obtain the age,

$$t_0 = \frac{2}{3H_\infty} \operatorname{Arcoth} \left(\frac{H_0}{H_\infty} \right) = 13.5 \times 10^9 \text{ year.} \quad (16.4)$$

We already calculated this age in a different but equivalent form in Exercise 15.7. See Exercises 16.2 and 16.3 also.

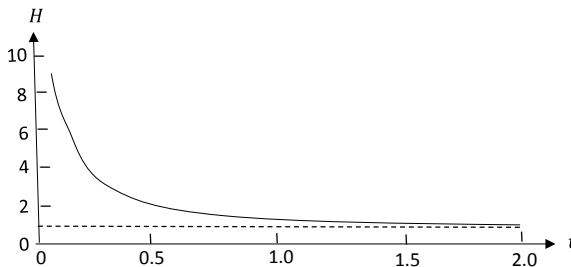


Fig. 16.1 The Hubble function for the flat LCDM universe. The function is shown in multiples of H_∞ and the time in multiples of $(3/2)H_\infty$, as in (16.2)

In Fig. 16.1 the Hubble function (16.2) is plotted, showing its explosive beginning at early times and its approach to the constant H_∞ at late times.

16.4 Transition Time for Matter to Dark Energy Dominance

After its first few hundred thousand years the universe was dominated by cold matter, meaning a cosmic matter fluid with negligible pressure. As it expanded the energy density of the matter decreased proportional to the inverse cube of the scale factor while the energy density of the dark energy, that is the cosmological constant, remained the same. We will calculate the time at which the two densities were equal, which we can call the transition time or the time of equality.

As usual we use a dimensionless scale factor that we set equal to unity at the present time. Then it is easy to obtain the value of the scale factor at the time of equality. Following the above comments and Sect. 14.4 we have for the evolution of the matter density and the dark energy density

$$\rho_m = \frac{\rho_{m0}}{a^3}, \quad \rho_V = \rho_{V0}, \quad (16.5)$$

where the subscript “0” refers, as usual, to the present time. Thus equality occurs when the scale factor is

$$a = \left(\frac{\rho_{m0}}{\rho_{V0}} \right)^{1/3} = \left(\frac{\Omega_{m0}}{\Omega_{V0}} \right)^{1/3}. \quad (16.6)$$

With presently measured values of about $\Omega_{m0} = 0.30$ and $\Omega_{V0} = 0.70$ this implies $a = 0.75$ and $z = 0.33$.

To determine the time of equality we use (16.1) but we write it in a form in which the present scale factor is explicitly equal to unity; that is

$$a(t) = \frac{[\sinh(\frac{3}{2}H_{\infty}t)]^{2/3}}{[\sinh(\frac{3}{2}H_{\infty}t_0)]^{2/3}}. \quad (16.7)$$

Equating (16.6) and (16.7) we obtain an equation for the time of equality t_e

$$\sinh\left(\frac{3}{2}H_{\infty}t_e\right) = \left(\frac{\Omega_{m0}}{\Omega_{v0}}\right)^{1/2} \sinh\left(\frac{3}{2}H_{\infty}t_0\right). \quad (16.8)$$

With the parameter values in Table 16.1 the numerical value of this is about $t_e = 9.9 \times 10^9$ year. Thus for about 3/4 of its existence the universe was dominated by cold matter.

16.5 Density Ratios and the Shape of the Universe

In the preceding sections we considered the flat LCDM universe, that is $k = 0$. Let us now relax that restriction and see how we might determine the sign and value of the curvature parameter k , which tells us the shape of the universe. Recall from Sect. 14.2 that one way to determine the sign of k was discussed in Chap. 14: if the total density of the universe exceeds the critical density in (14.7) then $k > 0$ and the universe is positively curved, finite and closed; if the total density is equal to the critical density then $k = 0$ and the universe is flat, infinite and open; if the density is less than the critical density then $k < 0$ and the universe is negatively curved, infinite and open.

In this section we will further study the behavior of the density ratios for matter and the vacuum in the LCDM universe, allowing for arbitrary curvature. This will also shed light on the effects of the cosmological constant or dark energy. As before we take the pressure to be negligible, which is well justified for times after a few hundred thousand years. That is, the matter is cold. The various ratios we obtain are surprisingly simple and of interest regarding observations. One of our final results is that present observations show that the universe is flat or almost flat—according to a reasonable definition of almost flat.

In this section we will explicitly write the present scale factor as $a(t_0) = a_0$ rather than take it to be 1 as we have usually done.

Let us begin with the density ratio for matter, $\Omega_m = \rho_m / \rho_{\text{crit}}$. The matter includes dark matter and ordinary baryonic matter. We wish to obtain an expression for this ratio as a function of the scale factor a so we can trace its behavior as the universe expands. For early times the result is particularly interesting. The cosmological equation (14.5a) gives us

$$\left(\frac{8\pi G}{c^4}\right)\rho_m = -\Lambda + 3\left[\frac{k}{a^2} + \frac{a'^2}{c^2 a^2}\right] = -\Lambda + \frac{3}{c^2 a^2}(a'^2 + kc^2). \quad (16.9)$$

From its definition in (14.7) the critical density obeys a similar equation

$$\begin{aligned} \left(\frac{8\pi G}{c^4}\right)\rho_{crit} &= \left(\frac{8\pi G}{c^4}\right)\left(\frac{3H^2c^2}{8\pi G}\right) = \frac{3}{c^2}H^2 = \frac{3}{c^2a'^2}(a'^2), \\ \rho_{crit} &\equiv \frac{3c^2H^2}{8\pi G}. \end{aligned} \quad (16.10)$$

Hence the matter density ratio at any time is given by

$$\Omega_m = \frac{\rho_m}{\rho_{crit}} = \frac{a'^2 + kc^2 - c^2a'^2\Lambda/3}{a'^2}. \quad (16.11)$$

In terms of the present density ratios from (14.19) this is

$$\begin{aligned} \Omega_m &= \frac{a'^2 - a_0^2H_0^2\Omega_{k0} - a^2H_0^2\Omega_{V0}}{a'^2}, \quad \Omega_{V0} \equiv \frac{\Lambda c^2}{3H_0^2}, \\ \Omega_{k0} &\equiv -\frac{kc^2}{a_0^2H_0^2}, \quad \Omega_{m0} \equiv \frac{8\pi G\rho_{m0}}{3c^2H_0^2}. \end{aligned} \quad (16.12)$$

We have repeated the definitions of the present values of the vacuum and curvature and matter ratios from (14.19) for convenience. But from (14.19) we may solve for a'^2 and thereby express this as

$$\Omega_m = \frac{\Omega_{m0}\left(\frac{a_0}{a}\right)^3}{\left[\Omega_{m0}\left(\frac{a_0}{a}\right)^3 + \Omega_{V0} + \Omega_{k0}\left(\frac{a_0}{a}\right)^2\right]}. \quad (16.13)$$

This is an elegant and informative relation. It tells us that for early times, when a is small, the matter density ratio must have been nearly one, quite independent of its present value. Similarly, for very large a in the future the matter density ratio must approach zero unless the cosmological constant is zero.

In similar manner we next calculate the ratio for the vacuum energy density or cosmological constant. From the definition of the vacuum energy density in (14.7) and the critical density noted above we have

$$\Omega_V = \frac{\rho_V}{\rho_{crit}} = \frac{\Lambda c^2}{3H^2} = \frac{c^2a'^2\Lambda}{3a'^2} = \frac{c^2a'^2\Lambda}{3a'^2} = \frac{a^2\Omega_{V0}H_0^2}{a'^2}. \quad (16.14)$$

Again we substitute for a'^2 from (14.19) to get

$$\Omega_V = \frac{\Omega_{V0}}{\left[\Omega_{m0} \left(\frac{a_0}{a} \right)^3 + \Omega_{V0} + \Omega_{k0} \left(\frac{a_0}{a} \right)^2 \right]}. \quad (16.15)$$

Notice that this ratio goes to zero at early times and unity for late times.

The ratio of vacuum energy density to matter energy density is, from (16.13) and (16.15),

$$\frac{\Omega_V}{\Omega_m} = \frac{\Omega_{V0}}{\Omega_{m0}} \left(\frac{a}{a_0} \right)^3. \quad (16.16)$$

This of course agrees with the results of Sect. 14.4: as the universe expands the vacuum energy density becomes more and more dominant.

The total density ratio of matter and vacuum energy is, from (16.13) and (16.15),

$$\begin{aligned} \Omega = \Omega_m + \Omega_V &= \frac{\Omega_{m0} \left(\frac{a_0}{a} \right)^3 + \Omega_{V0}}{\left[\Omega_{m0} \left(\frac{a_0}{a} \right)^3 + \Omega_{V0} + \Omega_{k0} \left(\frac{a_0}{a} \right)^2 \right]} \\ &= \left[1 + \frac{\Omega_{k0} \left(\frac{a_0}{a} \right)^2}{\left(\Omega_{m0} \left(\frac{a_0}{a} \right)^3 + \Omega_{V0} \right)} \right]^{-1}. \end{aligned} \quad (16.17)$$

It is this ratio which determines whether the universe is open or closed. This total energy density ratio, not including curvature, for LCDM is shown in Fig. 16.2. For $k = 0$ and $\Omega_{k0} = 0$ the ratio Ω is identically 1; for $k > 0$ and $\Omega_{k0} < 0$ it rises from 1 to a maximum and then decreases asymptotically to 1; for $k < -1$ and $\Omega_{k0} > 0$ it decreases from 1 to a minimum and then increases asymptotically back to 1. See Exercise 16.4 concerning the maximum and minimum values of the ratio.

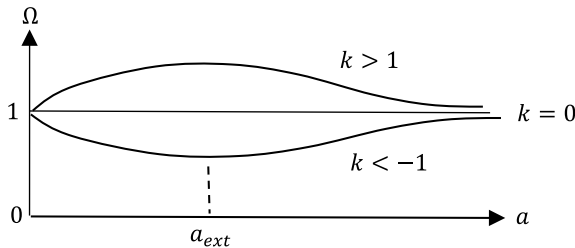


Fig. 16.2 Qualitative sketch of the total energy density ratio for the LCDM model. The extrema both occur at a_{ext} . See Exercise 16.4 for the value of a_{ext}

Consider for a moment the earliest times for which the LCDM model, neglecting radiation, could be roughly valid, which is for $z = a_0/a \approx 10^3$. At that time the vacuum energy density was negligible compared to the matter density. The present measured value for Ω_{k0} is consistent with zero but could be as large as about 10^{-2} . Thus from (16.17) the total density ratio Ω at the beginning of the LCDM era must have been quite close to unity, as is clear from

$$\Omega \cong 1 - \frac{\Omega_{k0} \left(\frac{a_0}{a}\right)^2}{\Omega_{m0} \left(\frac{a_0}{a}\right)^3} = 1 - \frac{\Omega_{k0}}{\Omega_{m0}} \left(\frac{a}{a_0}\right) \sim 1 \pm 10^{-5}. \quad (16.18)$$

Thus at the beginning of the matter era all the curves in Fig. 16.2 are quite close to 1.

Finally, recall from Chap. 13 that the curvature parameter k was first encountered as the inverse square of the radius of hypersphere. We therefore express k as the inverse square of a characteristic radius, $k \equiv 1/R_c^2$. From the definition of Ω_{k0} in (16.12) we can make a rough lower estimate of that radius by

$$\begin{aligned} \frac{c^2}{R_c^2} &= kc^2 = |\Omega_{k0}| a_0^2 H_0^2 = |\Omega_{k0}| \frac{1}{T_H^2}, \\ R_c &= \frac{cT_H}{\sqrt{|\Omega_{k0}|}} \sim 10^{11} \text{ ly} \quad \text{for } \Omega_{k0} \sim 10^{-2}. \end{aligned} \quad (16.19)$$

Thus the characteristic radius is at least of order 10 times the Hubble distance cT_H , so our observable universe lies well within it. This serves as a reasonable definition of “almost flat” (Adler 2005).

In summary, based on current observations it appears that the total energy density of the universe is close to critical, and the universe is spatially flat or nearly so. This is the currently favored theoretical case, the flat LCDM universe or standard model. However we again emphasize that cosmology is like all of science in that observations are the primary facts of life and are continually changing and improving.

16.6 Horizons and the Size of the Observable Universe

Let us next study how different events in the universe can influence each other. In special relativity this problem is easy since no influence can move faster than the speed of light: obviously we can be influenced only by events within our past light cone. In general relativity and cosmology the answer is similar and nearly as simple, but the past light cone in the spacetime of cosmology is just a little more subtle and interesting.

Recall that the motion of co-moving objects in the FLRW metric is very simple: they remain at coordinate rest, as indicated in Fig. 16.3. Since the scale factor

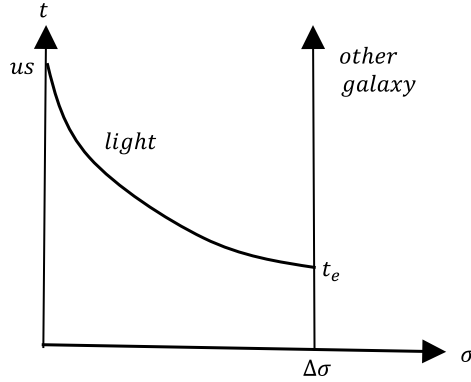


Fig. 16.3 Co-moving objects remain at coordinate rest in the FLRW metric. Light follows the indicated curve from the distant source to us, defining our past light cone

increases as time progresses the physical distance between two co-moving objects, such as galaxies, increases.

Let us ask how far we can see in the expanding universe? That is, what is the maximum distance, both coordinate and physical, that a source can be so that light from it has reached us? The question is very fundamental because it is equivalent to asking the size of the observable universe. Recall that light is characterized as having a null trajectory, $ds^2 = 0$. We impose this on the FLRW metric in the form (13.12b) to characterize the path of light so the coordinate distance differential obeys

$$ds^2 = c^2 dt^2 - a^2 d\sigma^2 = 0, \quad d\sigma = \frac{cdt}{a}. \quad (16.20)$$

To obtain the total coordinate distance σ we need to integrate this relation. Since the universe has been dominated by cold matter for most of its history, about 10 billion years, we do this using the scale factor for the period of matter dominance,

$$a = a_0 \left(\frac{t}{t_0} \right)^{2/3}, \quad a_0 = \text{present scale factor}, \quad t_0 = \text{present time}. \quad (16.21)$$

In this section we write explicitly the present value of the scale factor a_0 . From (16.20) and (16.21) we obtain the total coordinate distance $\Delta\sigma$ for light emitted at t_e and observed by us at t_0

$$d\sigma = \frac{cdt}{a_0} \left(\frac{t_0}{t} \right)^{2/3}, \quad \Delta\sigma = \frac{c}{a_0} \int_{t_e}^{t_0} dt \left(\frac{t_0}{t} \right)^{2/3} = \frac{3c}{a_0} (t_0 - t_0^{2/3} t_e^{1/3}). \quad (16.22)$$

For a source that emitted light near the beginning of the universe, $t_e = 0$, we thus have

$$\Delta\sigma = \frac{3c}{a_0}t_0 \equiv \sigma_{\text{hor}}. \quad (16.23)$$

Any source beyond this coordinate distance would not be observable since its light would not yet have reached us; σ_{hor} is our cosmic horizon, beyond which we cannot see. The coordinate horizon increases from a small value at early times to encompass more and more of the universe as time passes.

It is important to also calculate how far away, in physical distance, is a source now at the horizon. From (16.23) and the relation between coordinate and physical distances we obtain the physical distance ΔL ,

$$\Delta L = a_0\sigma_{\text{hor}} = 3ct_0. \quad (16.24)$$

That is, the source is three times as far away as the light from it has traveled to reach us!

The above result is remarkable and may be somewhat counter-intuitive. Since the scale factor approaches zero at very early times all the parts of the universe were then very close together. How is it then that light emitted from an object very near to our position has taken billions of years to reach us? It is because the expansion of the universe was initially so rapid that the matter outran the speed of light! This is evident from the fact that the Hubble function for the flat cold matter universe is $2/3t$ and diverges at early times. See also Example 15.1 for a discussion of recession velocity greater than c .

As we noted above the assumption of zero pressure and curvature and cosmological constant is reasonable for much of the history of our universe. The same sort of manipulations can be applied for the LCDM universe containing also dark energy but the analog of the integral in (16.22) does not reduce to a simple function like (16.23) for the horizon. Of course the relation nevertheless exists between the horizon and the present time: it is defined by the integral.

In Exercise 16.5 you are asked to obtain the horizon for a de Sitter universe using the same procedure as above. In the next section we will consider a different approach to the horizon for a flat de Sitter universe, that is one containing only dark energy.

16.7 Conformal Time

In the preceding chapters we have used the FLRW metric, which has $g_{00} = 1$ for a universal time coordinate. There is another type of metric which is often useful; it is one in which the line element is a multiple of the flat space metric of special relativity, so the behavior of light is essentially the same as in special relativity and light cones

are simple. It is called a conformal metric, as we have briefly noted previously. In this section we will show how such a metric can be obtained for the simple example of a flat de Sitter model universe. It will become clear that the same sort of manipulations can be used for other models, though not as simply.

We begin with the metric in standard flat FLRW form (13.12b) or (13.17), and expressed in Cartesian coordinates as

$$ds^2 = c^2 dt^2 - a(t)^2 d\vec{x}^2. \quad (16.25)$$

Our goal is to transform (16.25) into a form that is a multiple of the Lorentz metric of special relativity by using a new choice of time coordinate τ . That is, the metric expressed with this conformal time τ is to be

$$ds^2 = b(\tau)^2 [c^2 d\tau^2 - d\vec{x}^2], \quad \tau = F(t). \quad (16.26)$$

If we compare the metric forms (16.25) and (16.26) term by term we are led to the following relations

$$d\tau = F'(t)dt, \quad b(\tau)d\tau = dt, \quad b(\tau) = a(t). \quad (16.27)$$

From these relations we see that the new conformal time is given by an integral

$$F'(t) = \frac{1}{a(t)}, \quad \tau = F(t) = \int \frac{dt}{a(t)}. \quad (16.28)$$

These equations form the basis of the solution.

Let us apply the above analysis to the flat de Sitter model universe, which is believed to describe our real universe for very late times. As discussed previously in Chap. 13 the de Sitter scale factor is an exponential,

$$a(t) = a_i e^{\sqrt{\Lambda/3}ct}. \quad (16.29)$$

We choose the scale factor to be equal to 1 at the initial time $t = 0$ rather than the present time, so that $a_i = 1$. The conformal time is then given from (16.28) as

$$\tau = F(t) = \int_0^t dt e^{-\sqrt{\Lambda/3}ct} = \frac{1}{\sqrt{\Lambda/3}c} (1 - e^{-\sqrt{\Lambda/3}ct}). \quad (16.30)$$

This may be easily inverted to give t as a function of τ as in Exercise 16.8. However the most interesting quantity is the scale factor, which from (16.30) is

$$e^{-\sqrt{\Lambda/3}ct} = 1 - \sqrt{\frac{\Lambda}{3}}c\tau, \quad (16.31)$$

and thus from (16.27)

$$b(\tau) = a(t) = \frac{1}{1 - \sqrt{\Lambda/3}c\tau}. \quad (16.32)$$

The conformal metric is then

$$ds^2 = \left(\frac{1}{1 - \sqrt{\Lambda/3}c\tau} \right)^2 [c^2 d\tau^2 - d\vec{x}^2]. \quad (16.33)$$

The conformal time τ is given in terms of the FLRW time in (16.30); when the universe begins at $t = 0$ the conformal time is also $\tau = 0$. The end of the universe at $t = \infty$ corresponds to a finite conformal time

$$c\tau = \sqrt{3/\Lambda}, \text{ end of the universe!} \quad (16.34)$$

The line element (16.33) is singular at this final time, meaning that physical distances between comoving objects in the expanding universe are all infinite.

Figure 16.4 shows space–time in terms of the conformal time and Cartesian coordinates. As with the FLRW metric the world lines of co-moving galaxies are vertical lines, while the null rays of light are 45° lines. The past light cone of an observer at the end of the universe is specifically shown; from the figure it is clear that an object at a physical distance greater than $\sqrt{3/\Lambda}$ at $\tau = 0$ will never be seen by the observer!

The metric for a matter dominated universe can also be put into conformal form. This is left as an exercise for the reader; see Exercise 16.7. Unfortunately the conformal time and the form of the metric for the flat LCDM universe, the standard model, is not expressible in terms of elementary functions. For this and for any

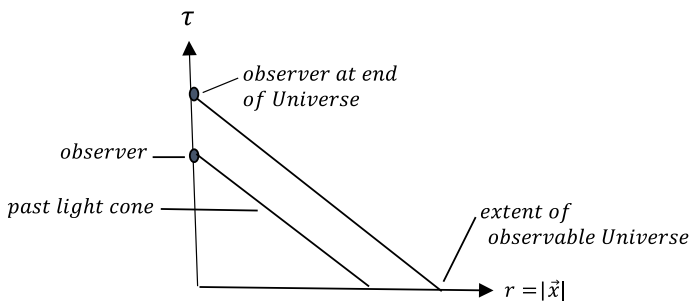


Fig. 16.4 Space–time in terms of Cartesian spatial coordinates and conformal time runs from $c\tau = 0$ to $c\tau = \sqrt{3/\Lambda}$ rather than infinity

scale factor (16.28) still serves as a definition of the conformal time in terms of the function $F(t)$.

Exercises

- 16.1 Use the relation (13.22) for the redshift and the dynamical equation (14.19) for the LCDM universe and show that the rate of change of z depends on the Hubble constant and the current matter density parameter according to

$$\frac{dz}{dt_0} = H_0 \left[(1+z) - \sqrt{\Omega_{m0}(3z + 3z^2 + z^3)} \right]$$

- 16.2 Show that the age of the universe according to (16.4) is equivalent to that obtained in Exercise 15.7.
- 16.3 For the LCDM universe use (14.19) to show that $\Omega_{V0} = H_\infty^2/H_0^2$. Then use this with (16.2) to express the age of the universe in terms of only Ω_{V0} and H_0 .
- 16.4 Consider the total density ratio obtained in (16.17). Show that the position of the extrema and the values of the density ratio are given by

$$\frac{a_{\text{ext}}}{a_0} = \left(\frac{\Omega_{m0}}{2\Omega_{V0}} \right)^{1/3}, \quad \Omega_{\text{ext}} = \left[1 + \frac{4\Omega_{k0}\Omega_{V0}}{\Omega_{m0}^2} \right]^{-1}.$$

Evaluate these for the parameter values in Table 16.1.

- 16.5 Following the procedure in Sect. 16.6 work out the horizon for the de Sitter universe. That is, obtain equations analogous to (16.23) and (16.24). You can do this for all three cases of the curvature k .
- 16.6 Suppose that for philosophical or esthetic reasons, you prefer the flat zero curvature model of the universe. Then you must be willing to contemplate an infinite real universe, which is conceptually problematic. Does the finite horizon and finite observable universe discussed in Sect. 16.6 make this more palatable? What of negative curvature?
- 16.7 Following the procedure of Sect. 16.7 put the cosmological metric in conformal form for the matter dominated universe with $k = 0$.
- 16.8 Invert (16.30) to give the standard cosmic time as a function of the conformal time. Also plot t versus τ as in (16.30).

Chapter 17

Earlier Times and Radiation



Abstract The CMB at present is very cold and has a very low energy density; however if we extrapolate back to earlier times, to a redshift of more than about 1000, we find that the temperature and energy density of the radiation was high enough that it was critically important in the evolution and behavior of the universe and its constituents. Indeed the atoms we observe today did not exist and a hot plasma dominated the universe until a few hundred thousand years. In this chapter we will study the scale factor and properties in the radiation era. The nearly perfect current isotropy of the CMB presents a theoretical problem and has led to the idea of inflation, wherein the universe underwent an extraordinary expansion in the very beginning. Moreover the very small anisotropies of the current CMB constitute essentially a photograph of the big bang which can give us a great deal of information about the earliest times.

17.1 Radiation and Temperature in Earlier Times

In the preceding chapters we considered a universe containing vacuum energy and matter at negligible pressure, as is well-justified for the present era. We only briefly mentioned earlier times, before galaxies were formed, and when the universe was hotter and radiation was important. Now we explicitly focus on such times with emphasis on the relative importance of matter and radiation to see explicitly how radiation becomes dominant.

Let us begin by considering the temperature and energy density of the cosmic microwave background (CMB) that fills today's universe, which we have already discussed in Chap. 13. The present CMB temperature is about 2.725 K. We can easily show that the temperature is inversely proportional to the scale factor. A well-known result from statistical mechanics and thermodynamics is that the energy density of black body radiation is proportional to the fourth power of its temperature, given by the Stefan-Boltzmann relation as

$$\rho_r = a_{\text{rad}} T^4, \quad a_{\text{rad}} = 5.6 \times 10^{-16} \text{J/K}^4 \text{m}^3. \quad (17.1)$$

But we also know from Sect. 14.4 that the radiation energy density is proportional to the inverse fourth power of the scale factor, as in (14.16). This relation and (17.1) tell us that the temperature of the radiation must be proportional to the inverse of the scale factor, or

$$\frac{T}{T_0} = \frac{a_0}{a}. \quad (17.2)$$

As usual the a_0 refers to the scale factor at the present time, which we often take to be equal to 1. This relation for the behavior of the temperature as the universe expands is remarkably simple. It has important consequences for the CMB spectrum and also for the behavior of the constituents of the universe at early times, long before the present LCDM universe. It clearly shows that the big bang was a hot big bang.

It should be emphasized that the temperature of the matter in the present universe is clearly not the same as that of the CMB since the two are not in thermal equilibrium. Clearly there is a great diversity of temperatures in the present universe, for example in the hot interior of stars and the cold of space. Equilibrium or lack of it is an interesting question for earlier times.

Having worked out the dependence of the temperature of the CMB on the scale factor let us consider how the spectrum of the CMB behaves during the expansion of the universe. Recall the Planck distribution law for black body radiation; it tells us that in a thermalized system at temperature T the number of photons in volume V , with frequency between ν and $\nu + d\nu$, is given by

$$dN = \frac{8\pi\nu^2 d\nu}{c^3(e^{h\nu/kT} - 1)} V. \quad (17.3)$$

Here h is Planck's constant and k is Boltzmann's constant. This famous distribution is sketched in Fig. 17.1.

The CMB spectrum was accurately measured by the cosmic background explorer satellite (COBE) and is in extremely good agreement with the Planck distribution, to about a part in 10^5 . This leaves little doubt that the CMB is indeed thermal radiation at 2.725 K (Fixsen 1993).



Fig. 17.1 Qualitative sketch of the Planck distribution, the spectrum of black body radiation

It is important to trace the evolution of the CMB spectrum during the expansion of the universe and verify that it does not change, that is (17.3) remains correct. To do this we consider the scaling of the various quantities in the Planck distribution as the scale factor expands to its present value of a_0 . The behavior of the physical volume is simple, as we have already mentioned in Chap. 16; as we look back in time the volume changes according to

$$V \rightarrow V_0 = \left(\frac{a_0}{a}\right)^3 V. \quad (17.4)$$

We also already know from Chap. 13 that the wavelength of a photon changes in proportional to the scale factor, and the frequency thus scales inversely proportional to the scale factor, so

$$\lambda \rightarrow \lambda_0 = \left(\frac{a_0}{a}\right)\lambda, \quad \text{and} \quad \nu \rightarrow \nu_0 = \left(\frac{a}{a_0}\right)\nu. \quad (17.5)$$

From (17.2) the temperature scales as

$$T \rightarrow T_0 = \left(\frac{a}{a_0}\right)T. \quad (17.6)$$

From the scaling relations (17.4)–(17.6) it is thus clear the Planck distribution does not change during expansion; that is

$$dN \rightarrow dN_0 = \frac{8\pi \nu_0^2 d\nu_0}{c^3(e^{h\nu_0/kT_0} - 1)} V_0 = \frac{8\pi \nu^2 d\nu}{c^3(e^{h\nu/kT} - 1)} V = dN. \quad (17.7)$$

Equivalently, we can say the scaling relations (17.4)–(17.7) are consistent.

There is now little doubt that the general picture of the CMB radiation being the remnants of the primordial big bang fireball is correct, due to its consistency and the excellent agreement with the black body spectrum. The general standard model scenario of the evolution of the universe during the LCDM era is also very likely to be correct; but we can also go much further back in time, into the radiation era, when the universe was filled with and dominated by radiation and hot plasma, quite near to time zero. Indeed our standard model of high energy particle physics allows us to understand with excellent confidence what happened as early as a second or so, and beyond that to about a microsecond with rather good confidence. These were the times when all the ordinary material of the present universe came into being, so this is an impressive claim. One of the reasons for such confidence is that some properties of the particles in the universe become simpler early in the radiation era due to the high temperature. For example, the hot quark gluon plasma prevalent at about a microsecond is probably well described as an ideal gas, since quarks and

gluons interact less strongly at high energies (Griffiths 1987; Peskin 2019). We will discuss the events of this era further, although briefly, in Chap. 18. Suffice it to say for now that our understanding of the radiation era is rather complete and dependable (Liddle 2003; Peebles 1993; Weinberg 1988).

Looking backward in time toward the radiation era let us ask for what value of the scale factor the temperature of the hot early universe dropped low enough that neutral atoms in their ground state could exist. Since most of the atoms in the universe are hydrogen this involves the atomic physics of hydrogen, which is well-understood. The term *recombination* refers to electrons binding with protons to form neutral hydrogen atoms, which are generally in a high energy state rather than the ground state. The excited neutral hydrogen atoms then emit photons and transition to the ground state; the photons can then interact with other hydrogen atoms. The term *decoupling* refers to the production of such photons that subsequently interact little with neutral hydrogen and propagate almost freely, often called free-streaming. These photons constitute the CMB which we observe today. Recombination and decoupling that occurred shortly afterward, are distinct but closely related events.

Note that recombination is a misnomer since the electrons and protons were never previously combined, but it is an established misnomer and almost universally used.

The time of decoupling thus corresponds to about the temperature at which hydrogen is largely ionized. This temperature can be estimated theoretically and measured experimentally, and is about 3000 K corresponding to an energy of 0.26 eV. Note that this is in the ballpark of the binding energy of hydrogen, 13.6 eV. See Exercise 17.1 and Liddle (2003). From this temperature and the present temperature of the CMB we can estimate from (17.2) the scale factor to be

$$\frac{a_{\text{dc}}}{a_0} = \frac{T_0}{T_{\text{dc}}} \approx \frac{2.725 \text{ K}}{3000 \text{ K}} \approx \frac{1}{1.1 \times 10^3}, \quad \text{dc denotes decoupling,} \quad (17.8)$$

so the redshift is about

$$z = \frac{a_0}{a_{\text{dc}}} - 1 = 1100. \quad (17.9)$$

The photons present at decoupling have been free-streaming ever since and are the ones we now see in the CMB; the decoupling event is also appropriately referred to as the *last scattering*. Thus we can think of the CMB as a photo of the big bang fireball redshifted in frequency by a factor of about 1000. As such we should expect it to contain a great deal of information about the universe at that time—and also earlier and later times. This is quite true as we will see in later chapters (Liddle 2003; Peebles 1993; Weinberg 1988).

In the following section we will study the scale factor in the radiation dominated era and use it to estimate the time at which decoupling occurred.

17.2 The Scale Factor and Basic Properties of the Radiation Era

Let us work out the scale factor for the radiation era and use it to estimate the time of decoupling and also the time when radiation and matter energy densities were equal. In Sect. 15.4 we obtained the scale factor for a universe dominated by cold matter. It is proportional to the $2/3$ power of the time, and we repeat it from (15.9) for convenience,

$$a = a_0 \left(\frac{t}{t_0} \right)^{2/3}. \quad (17.10)$$

It is also easy to obtain the scale factor for a universe dominated by radiation or a very hot gas. If we ignore all sources except radiation density in (15.3) we have

$$\int_0^a \frac{da}{a} \left[\Omega_{r0} \left(\frac{a_0}{a} \right)^4 \right]^{-1/2} = \frac{1}{2\sqrt{\Omega_{r0}}} \left(\frac{a}{a_0} \right)^2 = H_0 t. \quad (17.11)$$

We may write this in a form analogous to (17.10) as

$$a = a_0 \left(\frac{t}{t_0} \right)^{1/2}. \quad (17.12)$$

Thus the scale factor for radiation dominance is qualitatively similar to that for matter dominance; radiation involves a $1/2$ power whereas matter involves a $2/3$ power.

The complete scale factor for combined radiation and matter is also easy to calculate if we ignore the interaction between the two; this is not likely to be a very good approximation since the matter was charged so it should not be expected to be very accurate but it is an interesting theoretical exercise. (Note that electrons and matter move slowly enough to be considered “cold” as far back as when $kT \sim 0.5$ MeV; see Exercise 17.3.) We only need to evaluate the integral in (15.3) with matter and radiation constituents. The integral and its evaluation are

$$\begin{aligned} H_0 t &= \int_0^a \frac{da}{a} \left[\Omega_{m0} \left(\frac{a_0}{a} \right)^3 + \Omega_{r0} \left(\frac{a_0}{a} \right)^4 \right]^{-1/2}, \\ t &= \frac{2}{3} \frac{1}{\sqrt{\Omega_{m0}}} \left[\left(\frac{a}{a_0} - 2\varepsilon \right) \sqrt{\frac{a}{a_0} + \varepsilon} + 2\varepsilon^{3/2} \right] T_H, \\ \varepsilon &\equiv \frac{\Omega_{r0}}{\Omega_{m0}}, \quad T_H \equiv \frac{1}{H_0}. \end{aligned} \quad (17.13)$$

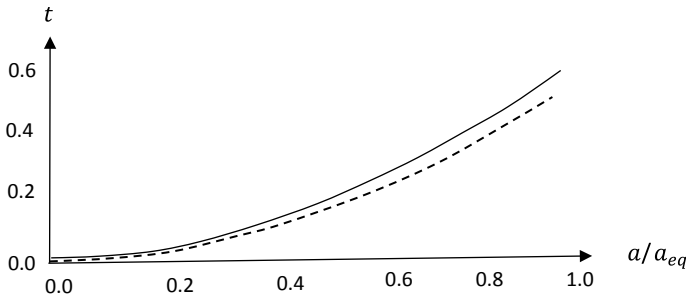


Fig. 17.2 The scale factor is in units of ε , which is its value at the time when matter and radiation energy densities are equal

The present ratio of radiation to matter energy defined in (17.13) is small, about $\varepsilon = 1.27 \times 10^{-4}$. It is straightforward to check that the limit of (17.13) for large a gives the scale factor for matter dominance in (17.10) and for small a gives the scale factor for radiation dominance in (17.12). See also Exercise 17.2. A plot of the scale factor (17.13) is shown in Fig. 17.2 along with the radiation-only curve according to (17.12).

Equation (17.13) allows us to calculate the time of decoupling, which we discussed in Sect. 17.1. In (17.8) we found that decoupling occurred for about $a_0/a = 1100$, based on the temperature at which hydrogen is ionized. Using the values in Table 16.1 for the parameters in (17.13) we find for the decoupling time

$$t_{\text{dc}} \cong 4.1 \times 10^5 \text{ year.} \quad (17.14)$$

A more detailed analysis gives about 380,000 year, so our estimate is not too bad (Peebles 1968; Smoot 2006). See Exercise 17.4 for other rough estimates.

Another time of interest is that for which the energy density in radiation decreased to become equal to that in matter. Since the energy density in matter is proportional to the inverse cube of the scale factor and that in radiation is proportional to the inverse fourth power equality occurs when $a/a_0 = \Omega_{r0}/\Omega_{m0} = \varepsilon$. Then with the parameter values in Table 16.1 we find the time for equality to be

$$t_{\text{eq}} = \frac{2}{3} \frac{\varepsilon^{3/2}}{\sqrt{\Omega_{m0}}} [2 - \sqrt{2}] T_{\text{H}} \cong 1.4 \times 10^4 \text{ year.} \quad (17.15)$$

Thus the time of equality is earlier than the time of decoupling in (17.14).

We stress that the above numbers for the time of decoupling and radiation matter equality are only quite rough estimates since they ignore interactions between the matter and radiation (Peebles 1968).

17.3 The Isotropic CMB and the Horizon Puzzle

As we have noted the CMB is extremely uniform and has a very precise black body spectrum; its temperature varies by only about a part in 10^5 over all directions of the sky. The very small variations have turned out to be very important as a probe of the early universe, as we will presently discuss. However the uniformity presents a problem: it tells us that the big bang fireball at decoupling time had nearly the same temperature everywhere. When we encounter in the laboratory a system such as a container of water with a very uniform temperature we naturally expect that it has achieved equilibrium over a substantial period of time and the uniform temperature is the result of the increase of entropy. But we can show that no such explanation can hold for the fireball in the context of the cosmological theory we have developed so far; this is because distant parts of the fireball could not have influenced each other before decoupling due to the finite speed of light and the rapid expansion of the universe entailed by the scale factor obtained in Sect. 17.2.

Explicitly we will show that according to the theory as developed so far we should not expect two regions of the sky to have precisely the same CMB temperature if they are more than a small number of degrees apart. Consider two sources of the cosmic background radiation, that is the big bang fireball, at coordinate distance σ_0 from us the observers, and σ_{12} coordinate distance from each other, subtending an angle θ , as shown in Fig. 17.3.

Source 1 could influence source 2 at the time of decoupling only if it is within the past light cone of source 2. We earlier studied this kind of problem in Sect. 16.6, but applied to the matter dominated era. Light going from source 1 to source 2 follows a null geodesic so

$$ds^2 = c^2 dt^2 - a^2 d\sigma^2 = 0, \quad d\sigma = \frac{cdt}{a}. \quad (17.16)$$

Thus the coordinate distance between sources 1 and 2 for radiation emitted at time zero and propagating until decoupling is

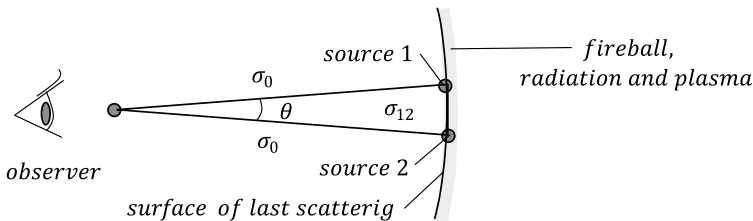


Fig. 17.3 Two sources of cosmic background radiation as seen by us, separated by an angle θ . The distances are all coordinate distances between co-moving objects and do not change with time

$$\sigma_{12} = c \int_0^{t_{\text{dc}}} \frac{a_0}{a(t)} dt = c \int_0^{t_{\text{dc}}} dt \left(\frac{t_0}{t} \right)^{2/3} = 3c(t_0^2 t_{\text{dc}})^{1/3}. \quad (17.17)$$

We are only interested in rough answers so the scale factor for matter used in (17.17) is adequate. Sources separated by more than this coordinate distance could not have influenced each other.

Exactly the same analysis for light traveling from the two sources to us during the matter era gives the coordinate distance to be

$$\sigma_0 = c \int_{t_{\text{dc}}}^{t_0} dt \left(\frac{t_0}{t} \right)^{2/3} \cong c \int_0^{t_0} dt \left(\frac{t_0}{t} \right)^{2/3} = 3ct_0. \quad (17.18)$$

We also obtained this σ_0 in (16.24). Thus the maximum angle of separation that we should expect to see between regions of precisely the same temperature is very roughly

$$\theta = \frac{\sigma_{12}}{\sigma_0} = \left(\frac{t_{\text{dc}}}{t_0} \right)^{1/3}. \quad (17.19)$$

Since the time of decoupling is 4×10^5 years and the age of the universe is of order 1.4×10^{10} years this angle is about 0.03 radians or a few degrees.

But we observe the CMB temperature over the whole sky to be quite isotropic, to about 10^{-5} , so there is a puzzle as to why regions outside the maximum causal angle (17.19) could be in such nearly perfect thermal equilibrium with each other: they could never have been in causal contact or influenced each other before decoupling according to the theory discussed up to this point. This is known as the horizon or isotropy puzzle, and brings our understanding of the early universe into question. However by looking at the puzzle from the opposite point of view we can take it as a clue to the nature of the universe before the radiation era. The horizon puzzle is one of the main motivations for the concept and theory of inflation which we will study in Chap. 19 (Peebles 1993; Liddle 2003). Exercises 17.5 and 17.6 give an indication of how inflation might solve the horizon puzzle.

17.4 The Anisotropies of the CMB

While the CMB spectrum is that of a black body and isotropic to high accuracy, there are tiny deviations predicted by theory and verified by observations. Because of this the detailed CMB spectrum has become an important tool in cosmology since it depends on all the contents of the universe and their behavior during the radiation

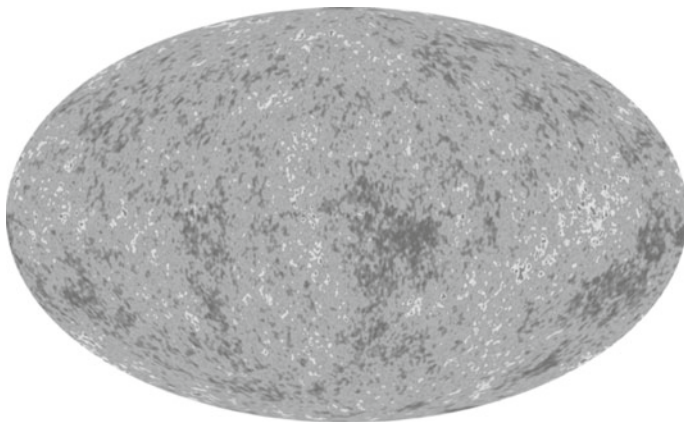


Fig. 17.4 An illustrative image of the CMB by the WMAP satellite. The lighter regions are slightly hotter than the darker regions

era and even earlier times when no ordinary matter even existed. We will discuss some observational aspects of this problem in this section and the theoretical aspects further in Chap. 19 (Peebles 1965).

Observations of the CMB by the COBE, WMAP and PLANCK satellites can be viewed as pictures of the big bang fireball at the time of decoupling when its temperature was about 3000 K at a redshift of about 1100 (NASA 2019). The structure in the pictures is due to temperature variations, and an example from the WMAP satellite is shown in Fig. 17.4. The lighter regions are hotter.

The standard way to analyze such CMB images is to express the temperature T as a function of the spherical coordinate angles θ, φ and expand fractional differences in spherical harmonics $Y_{\ell m}$ according to

$$\frac{T(\theta, \varphi) - \bar{T}}{\bar{T}} = \frac{\Delta T}{\bar{T}} = \sum_{\ell m} a_{\ell m} Y_{\ell m}, \quad m = -\ell \text{ to } \ell, \quad (17.20)$$

where $\bar{T}$ is the average temperature and the $a_{\ell m}$ are the expansion coefficients. This is the spherical analog of taking a Fourier transform with Cartesian coordinates and the $a_{\ell m}$ are the analogs of the Fourier transforms; this is a standard approach in wave analysis. The quantity of interest is the *angular power spectrum*, the average of the $a_{\ell m}$ over the m values,

$$C_\ell = \langle |a_{\ell m}|^2 \rangle. \quad (17.21)$$

It is this quantity that is often displayed in graphs and compared to theoretical predictions.

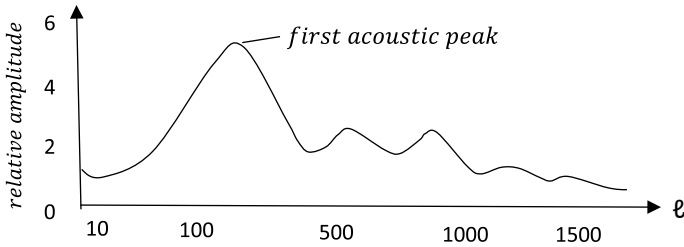


Fig. 17.5 A sketch of the CMB power spectrum obtained from an LCDM model

Calculation of the C_ℓ power spectrum from theory is an interesting but formidable task; it requires consideration of the physics of the materials in the universe during the radiation era, which includes such things as photon electron interactions, the type and density of neutrinos, etc. It also requires analysis of physics before and after the radiation era. Most important it requires an understanding of the wavelength of density fluctuations due to standing sound waves, called baryon acoustic oscillations or BAOs (Knox 2019). We will give a brief conceptual discussion of some of these issues in Chap. 19.

Because of the complexity of the problem the calculation of the CMB spectrum is usually done using openly accessible computer programs that turn cosmological models into spectra quite rapidly; one useful reference is Tegmark (2019). An illustrative example of such a theoretical spectrum is shown in Fig. 17.5.

One important aspect of comparing theory and CMB observation is to ascribe specific features of the spectrum to various physical causes. The spacing of the prominent peaks in Fig. 17.5 is particularly interesting; it allows us to measure distances around the time of emission of the CMB, and from that estimate the Hubble constant. We can see intuitively how this can be done: the peaks are due to standing sound waves, the BAOs, at and before the time of emission; the waves correspond to regions of high and low density. Combined with a theoretical understanding of the speed of sound at that time this provides a distance scale, and that distance scale acts as a meter stick in the sky (Knox 2019).

Thus, as we mentioned previously in Sect. 16.1 and Appendix 1 in Chap. 13 the comparison of theory with the observed CMB spectrum provides a measurement of important cosmological parameters. Indeed it is fair to say that the study of the CMB spectrum is now one of the most important tools in early universe cosmology (NASA 2019).

Exercises

- 17.1 Why is the temperature 0.26 eV at which hydrogen atoms in the early universe were largely ionized considerably less than the binding energy 13.6 eV? Give a rough qualitative answer. It is possible to estimate the temperature accurately as in Liddle (2003) and Peebles (1993).

- 17.2 Show that the limit of (17.13) for small values of the scale factor is the same as (17.12) for a pure radiation filled universe. Show that the limit for large values of the scale factor is the same as (17.10) the pure matter filled universe.
- 17.3 What is the kinetic energy of an electron in thermal equilibrium with radiation at the time of decoupling? What is its velocity as a fraction of the speed of light? Is the gas of electrons hot or cold?
- 17.4 Estimate the time of decoupling as we did in Sect. 17.2 but use the assumption that the scale factor is that for matter given in (17.10). Repeat for the pure radiation scale factor in (17.12). Compare the results.
- 17.5 Repeat the analysis of Sect. 17.3 if the scale factor of the very early universe is that of flat de Sitter space with zero curvature. Note that the universe for this case could begin at any negative time. Obtain the analog of (17.19). Is there a horizon puzzle for this choice of scale factor? We will return to this problem in Chap. 19.
- 17.6 Repeat Exercise 17.5 for the cases of positive and negative curvature parameter k . Is there a horizon puzzle for these cases?
- 17.7 Show that features in the CMB with an angular size of about $\Delta\theta$ will show up in the power spectrum at values of about $\ell = \pi/\Delta\theta$.

Chapter 18

A Brief Historical Overview of the Universe



Abstract We may use our theories and presently observed properties of the universe to run the cosmic clock backwards and view the history of the universe in reverse. This takes us from the present universe of stars and galaxies, to the radiation era of hot plasma when the presently observed matter formed, and back to the era commonly described by inflation. We may finally run the clock all the way back to a putative Planck era when spacetime itself was likely fundamentally different from what it is at present, and for which we have no accepted theory.

18.1 Overview

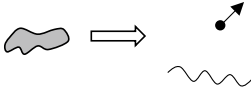
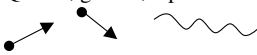
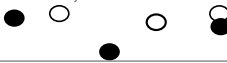
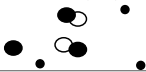
This chapter is a very brief and superficial account of the contents, temperature and size of the universe during its history (Freedman 2006). We have already traced the history from the present back into the radiation era in the preceding chapters. As we noted in Sect. 17.1 we can track and understand the history with some confidence back to about a microsecond, which is a rather remarkable claim (Peebles 1993; Weinberg 1988).

We now live in a universe of stars and galaxies made of atoms, ionized plasma, and a great deal of dark matter of unknown nature. The vacuum also appears to have a nonzero energy density, although we do not understand why it has the magnitude it does; the vacuum or dark energy density is the dominant constituent in the universe as we discussed in previous chapters. We are also bathed in the CMB radiation sea, and almost certainly a similar sea of neutrinos, which has yet to be detected. As we discussed in Chap. 17 the energy density in matter is about 10,000 times the energy density in the CMB radiation at present, but we know the radiation density was much greater in the radiation era. The young universe was hot and had a different composition than at present as we discussed in the preceding chapter. In this chapter we will run the cosmic clock of the universe further backwards to even earlier times, and thus a smaller scale factor and higher temperature; we want to see what the important contents of the universe were, based on the physics of atoms, nuclei, nucleons, and quarks and leptons.

The present universe contains mainly hydrogen atoms, about 90% by number, and Helium atoms, about 10% by number, with everything else being much less. Only hydrogen and helium, and a small amount of deuterium and helium 3 and lithium, have existed during most of the life of the universe, and we call these the *primordial* elements. The heavier elements that are familiar and abundant on earth were all formed millions and billions of years later by fusion in stellar interiors, and subsequently ejected into space by supernova explosions; many of the heaviest elements are now believed to be formed in kilonovas, the collisions of neutron stars with other neutron stars or with black holes, indicated by gravitational waves (Peebles 1993; Kasen 2017; Holz 2018). Thus the material that we and our planet are mainly composed of had quite a spectacular origin in diverse stellar explosions. However we may ignore elements other than hydrogen and helium in much of our cosmological discussion.

Table 18.1 shows major events and contents of the universe; obviously everything is quite approximate and the sketches of the contents are obviously symbolic.

Table 18.1 The life and contents of the universe in brief

Time	Temperature/energy	Major event	Notable contents
0?	∞ ?	Chaos?	Chaos? 
10^{-43} s	10^{19} GeV	Spacetime begins	Quantum geometry? 
10^{-35} s	10^{17} GeV	Rapid expansion	Inflaton field 
10^{-34} s	10^{16} GeV	Matter appears	Diverse particles 
10^{-6} s	1 GeV	Nucleons condense	Quarks, gluons, leptons 
1 s	1 MeV	Nuclei condense	Protons, neutrons 
10^5 year	1 eV	Atoms condense	Nuclei H, D, He, electrons 
10^8 year	60 K	Stars, galaxies condense	Hydrogen atoms 
10^{10} year	2.7 K	GR discovered	Stars, planets, physicists 

18.2 Condensation of Stars and Galaxies

Stars and galaxies began to form early in the history of the universe, about 200 million years after the end of the radiation era, and they are still being formed. A star forms when a large enough ball of gas with a density excess begins to contract due to gravity. As the ball contracts it heats up, and eventually the temperature reaches the point at which thermonuclear fusion begins in the gas (Chandrasekhar 1939; Misner 1973). The dominant overall nuclear fusion reaction is $4p \rightarrow He + 2e^+$ with several neutrinos also being emitted. Energy released by fusion stops the gravitational contraction by increasing the pressure due to heat and by direct radiation pressure. The star then becomes a stable energy emitting member of the main sequence for billions of years. Finally, enough of the fusible elements are used up that the star dies, sometimes in a supernova explosion. We will not go into such stellar astrophysics more deeply here since it is a major subject in itself and we can proceed with a sketchy understanding for our study of cosmology (Carroll 2017; Dar 2006).

A galaxy forms when billions of stars combine into a large system. Galactic evolution is an active field of research with many unknowns. Oddly enough we seem to know more about quarks and the composition of nucleons than about galaxies (Dar 2006; Quigg 2006).

We may think of stars forming from gas clouds by instability to gravitational attraction as a sort of stellar condensation process. We may similarly think of galaxies as forming from stars and gas as a sort of galactic condensation process.

18.3 Condensation of Atoms

Let us again run the cosmic clock backwards. As we have discussed in Chap. 17 the universe grows hotter until a temperature of about 3000 K (roughly $kT = 0.3$ eV) is reached, which happens at about 10^5 years or 10^{12} s after the big bang. Before this recombination time atoms cannot exist because radiation and collisions ionize them, so the universe is largely composed of a plasma of hydrogen and helium nuclei (protons and alpha particles) and electrons along with many photons and neutrinos. We can view this process as atoms condensing from the hot plasma. The plasma is of course charged and thus opaque to light and its contents (except perhaps the neutrinos) are in thermal equilibrium due to electromagnetic interactions between charged particles (Peebles 1993). Early on when the mixture is very hot it is fairly well described by an ideal gas equation of state with $p = \rho/3$.

Note that neutrinos produced in early times interact very weakly and should still exist now in the form of a neutrino background, analogous to the CMB radiation, but the neutrino sea has not yet been detected. The role of neutrinos in cosmology is presently under intense study; it depends on their mass, which is small but nonzero (Dvorkin 2019).

18.4 Condensation of Nuclei

Again we run the clock backwards, well before the 10^{12} s decoupling time, to a time when the thermal energy is about 1 meV. At higher temperature nuclei cannot exist because they are disintegrated by collisions and γ rays in the radiation. Thus we can think of nuclei as condensing from a plasma gas of nucleons at this time, about 1 s. The process of nuclear formation is commonly called nucleosynthesis. The plasma before this time is composed mainly of neutrons and protons and electrons and many photons and neutrinos (Weinberg 1988).

The theory of primordial nucleosynthesis has been very successful in predicting the abundance of the primordial light elements. For example, the number ratio of helium to hydrogen atoms is predicted to be about 1/10, in good agreement with observation. This calculation assumes that essentially all the neutrons become bound in helium nuclei, and thus it depends critically on the relative abundance r of neutrons and protons just before the helium nuclei condense. To obtain this ratio requires analysis of the beta decay reaction during cooling of the nucleon gas,

$$p + \bar{\nu} \rightarrow n + e^+. \quad (18.1)$$

The result is that the ratio was about $r = 0.17$ when neutrinos decoupled, and that this dropped to about $r = 0.14$ due to neutron decay by the time the helium condensed. It is easy to see that the abundance of helium follows as

$$\frac{\text{He}}{\text{H}} = \frac{r/2}{1-r} = 0.08 \quad (18.2)$$

as we anticipated.

The predicted abundance of deuterium (heavy hydrogen) is about 10^{-4} , that of helium 3 (light helium) is about 10^{-5} , and that of lithium is about 10^{-10} ; all are consistent with observation (Wagoner 1967). These predicted abundances are sensitive to the composition and temperature of the universe at the time of nuclear condensation, and this places important constraints upon these properties (Freedman 1967). One interesting result is that the abundance of ordinary nucleonic matter must be about an order of magnitude less than the present critical density, and thus that the dark matter is unlikely to be ordinary nuclear matter (Weinberg 1988).

18.5 Condensation of Nucleons

Yet again we run the clock backwards, well beyond the 1 s time of nuclear condensation, to a time when the temperature energy is about 1 GeV. At higher temperature even nucleons are not stable, but decompose into quarks and gluons. Thus we can

think of nucleons condensing from a quark gluon plasma at this time, about 10^{-6} s. The stuff of the universe before this time is a dense hot plasma of quarks and gluons and leptons (e and μ and τ and their neutrinos) and many photons (Quigg 2006).

High energy particle experimentalists are now trying to create quark gluon plasmas at accelerator laboratories by colliding heavy nuclei. The results are interesting for their relevance to early universe physics.

A question under current debate is why the universe appears to be composed almost entirely of matter and very little anti-matter. One might naively expect a mixture of roughly 50% matter and 50% anti-matter, but this is definitely not observed (Moskowitz 2019; Peebles 1993).

18.6 Inflation

Recall the horizon or isotropy puzzle, which we discussed in Sect. 17.3: it is difficult to see how the universe could have been as homogeneous and isotropic as it appears to have been at the time of decoupling. But also recall that we noted in Exercise 17.5 that the exponentially expanding de Sitter model universe has no horizon. Many theorists believe that a period of very rapid expansion of the universe, described roughly by the de Sitter model, provides the best resolution of the horizon puzzle; this very rapid expansion is called inflation. It is postulated to have occurred well before the quark gluon plasma era, say at about 10^{-36} s. There are many versions of inflation theory, and it is better to consider it a general scenario rather than a single theory. Most versions postulate a scalar field called the inflaton as the dominant ingredient of the universe. It is now the most widely accepted solution of the horizon puzzle and we will discuss it in Chap. 19. It is also relevant to the spectrum of the CMB radiation, which provides an observational test (Freedman 2006; Linde 2007).

The end of the inflation era is widely called reheating, at which time the particles and fields that later filled the universe were somehow formed after inflation. There is little observational information concerning this transition, but much theoretical speculation (Kofman 1996).

18.7 Planck Era

Finally we run the clock backwards for the final time to such high temperature and energy that we simply do not know what happens but can only make informed guesses. Perhaps interesting and strange things happen when the temperature approaches $kT = 10^{17}$ GeV, at which time it is believed by many theorists that the strong, electromagnetic and weak forces become equal. There are many possibilities.

However, one thing seems to be clear about very early times: at about $kT = 10^{19}$ GeV and 10^{-43} s the description of gravity by general relativity is no

longer valid since we must take account of quantum effects. Indeed, according to general ideas of quantum theory spacetime itself undergoes large quantum fluctuations and is not describable in classical terms (Adler 2010). This time period is called the Planck era since Planck was the first to realize that the fundamental constants h , c , G define a natural scale; this very small scale is likely to be the relevant one for this earliest phase of the universe.

At this point we must finally give up on running the cosmic clock backwards, since we do not have a believable quantum theory of gravity. We likely do not even know if times earlier than the Planck time have meaning. We will discuss some basic concepts and speculations concerning quantum theory and gravity and the Planck scale in Chap. 19.

Exercises

There are none of the usual exercises for this descriptive and qualitative chapter. The reader interested in more information may read Weinberg's well-known little book *The First Three Minutes*, which is clear and informative but a little out of date (Weinberg 1988). Another source of more recent information is Wikipedia; it contains much information, but since it is not peer reviewed there is no guarantee of its accuracy. More dependable discussions are in the readable undergraduate text by Liddle (2003) and the authoritative text by Peebles (1993).

Chapter 19

Inflation and Some Questions



Abstract Inflation involves an extremely rapid expansion of the universe, which can be qualitatively described in terms of a de Sitter model universe. We will briefly discuss the dominant current view of inflation, which is based on one (or several) scalar fields that caused the early universe to expand exponentially much like a de Sitter universe. Quantum fluctuations during the expansion may be used to explain the subsequent structure we observe on the cosmological scale. Observational data on the inflationary era however are sparse and many questions remain. Other questions of note that we will briefly mention are the physical nature of dark matter and dark energy, and the quantum properties of the universe in the earliest times.

19.1 Basic Ideas of Inflation

We saw in Sect. 17.3 that there is a puzzle as to how the universe could have been so isotropic at the decoupling time. The cosmic background radiation is observed to be isotropic to about 10^{-5} over the whole sky, whereas the models we have discussed predict that only a few degrees of the sky could have been causally connected and thermalized at the time of the last scattering of the CMB. One might simply postulate that the universe was initially very isotropic, but that is not a satisfying answer. One favored approach to a solution is to appeal to a scenario called inflation, wherein there is postulated to be a period of extremely rapid expansion before the radiation era. This produces a horizon such that the entire observable universe was once in causal contact and thus could have been thermalized (Freedman 2006; Peebles 1993; Linde 2007; Liddle 2003).

At a conceptual level the solution to the horizon puzzle is quite simple. Instead of the situation shown in Fig. 17.3 in which the coordinate distance σ_{12} is small and the angle θ is thus also small we ask instead that σ_{12} be quite large so the coordinate horizon distance is large enough to encompass our entire view of the big bang fireball. This obviously means that the scale factor a must be such that the integral in (17.17) for σ_{12} is larger than the coordinate distance from the source to us in (17.18). It is also obvious that there are many ways we could choose a so that this is true; indeed it is easy to choose the scale factor a so the integral is divergent and σ_{12} is infinite.

Thus the demand we wish to place on the scale factor is that the horizon coordinate distance during inflation be large, or perhaps infinite. That is, explicitly

$$\sigma_{12} = c \int_{t_I}^{t_r} \frac{a_0}{a(t)} dt > \sigma_0. \quad (19.1)$$

Here t_I denotes the beginning of the inflation era and t_r denotes the end of the inflation era and beginning of the radiation era. The horizon at decoupling will be larger than this and thus large enough to resolve the horizon puzzle. In the following paragraphs we will elucidate how this can come about.

As a simple example of how to satisfy the demand in (19.1) consider an ad hoc choice of a scale factor, a power of the cosmic time,

$$a = a_0 \left(\frac{t}{t_0} \right)^m, \quad (19.2)$$

where the power m is taken to be large, say of order 10 or more. Then the integral for σ_{12} in (19.1) diverges and the horizon is infinite. This model of inflation is naturally called power law inflation; it corresponds to a model universe filled with a fluid with an unusual equation of state. The power law model of inflation provides an excellent pedagogic example, but it is not presently a favored model, so we have relegated further discussion to Appendix 1 (Peebles 1993).

As a second example consider another ad hoc choice, a de Sitter space with an exponential scale factor, which we have already discussed in Chap. 13 and Exercise 17.5. Recall that de Sitter space describes a universe containing only vacuum energy, or equivalently a cosmological constant. It is very important to distinguish this vacuum energy during inflation from the present vacuum energy in the LCDM model; the energy scale during inflation is vastly larger. This model provides an illuminating phenomenology for the basic features of inflation. We will use it to work out the coordinate horizon σ_{12} and its relation to the coordinate distance σ_0 of the last scattering. Thus we choose during inflation

$$a = Ae^{H_I t}, \quad H_I = \text{const.} \quad (19.3)$$

A qualitative sketch is shown in Fig. 19.1. To make a rough estimate of the constant A we equate this to the scale factor for the radiation era at its beginning as given in (17.12) and thereby obtain.

$$a = a_0 \left(\frac{t_r}{t_0} \right)^{1/2} e^{H_I(t-t_r)}, \quad t \leq t_r, \quad t_0 \text{ present time.} \quad (19.4)$$

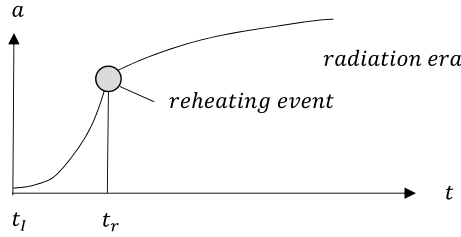


Fig. 19.1 The scale factor during inflation, the reheating event marking the end of inflation, and the scale factor in the early radiation era

Then the horizon at the end of inflation is a simple integral from (19.1),

$$\sigma_{12} = \frac{c}{H_I} \left(\frac{t_0}{t_r} \right)^{1/2} [e^{H_I(t_r - t_I)} - 1] \cong \frac{c}{H_I} \left(\frac{t_0}{t_r} \right)^{1/2} e^{H_I(t_r - t_I)}, \quad (19.5)$$

where we have assumed that the exponential in the parenthesis is much larger than 1. If the horizon puzzle is to be solved this σ_{12} must be larger than the coordinate distance σ_0 from us to the last scattering surface. We calculated this in (17.18) and repeat it here,

$$\sigma_0 = 3ct_0. \quad (19.6)$$

After some rearrangement we then may write the ratio needed from (19.1) in the form

$$\frac{\sigma_{12}}{\sigma_0} = \frac{1}{3H_I} \left(\frac{1}{t_0 t_r} \right)^{1/2} e^{H_I(t_r - t_I)} = \frac{1}{3} \left(\frac{t_r}{t_0} \right)^{1/2} \frac{e^{H_I t_r}}{H_I t_r} > 1. \quad (19.7)$$

where we have assumed in the last expression that $t_r \gg t_I$.

As yet there is no clear observational evidence on the time and energy scale for inflation, but there are many speculations by various theorists (Peebles 1993; Linde 2007). For illustration we will suppose inflation runs from zero to about 10^{-33} s and refer the reader to the references for more information. Then we may turn (19.7) into an equation for the quantity $H_I t_r = N$, which is the number of e-foldings during the inflation era. With order of magnitude estimates for the various times in (19.7) we then find the demand on N to be

$$\frac{1}{N} e^N > 3 \left(\frac{t_0}{t_r} \right)^{1/2} \sim 10^{22}, \quad \text{for } t_0 \sim 10^{10} \text{ year}, \quad t_r \sim 10^{-33} \text{ s}. \quad (19.8)$$

Numerically this is satisfied for N of order 50 or 60, which implies a huge expansion of order 10^{24} in a very short time, the characteristics of inflation.

In summary, if the universe expands exponentially at a very early time, and the exponential expansion continues for about 50–60 e-foldings, then the horizon is such that all points in the sky were once in causal contact and could have been thermalized to the same temperature: in terms of this phenomenological picture the horizon puzzle may be thereby solved.

Presumably the era of inflation ends when its large associated energy somehow is transformed into the particles of matter that we see in the universe today; this process is termed reheating as we have noted in Chap. 18; clearly it is not a very descriptive phrase since there was probably no previous heating. The mechanism for reheating is a subject of theoretical study, but little is known (Peebles 1993).

In the next section we will discuss a flexible and popular mechanism to describe inflation using field theory.

19.2 Inflation Via Scalar Fields

Most presently favored theoretical models for inflation behave qualitatively similar to the above phenomenological example using de Sitter space. The vacuum energy density in such a de Sitter space can be thought of as a constant density fluid with the equation of state $p = -\rho$. It is possible to devise a scalar field which behaves much like such a fluid, and this gives a plausible and flexible type of theory for the material and geometry during the inflationary era. Indeed there are many versions of such fields and models, and naturally the fields are generically called inflaton fields. In this section we will outline one of the simplest examples of an inflaton field, but there is much current research in this field and we will only give a general discussion (Liddle 2003; Peebles 1993; Linde 2003).

We will assume the reader has some familiarity with the Lagrangian approach to field theory; for the reader who is not familiar with field theory. Appendix 2 gives a brief overview. It is now almost universal practice in field theory to use “natural units” in which $\hbar = 1$ and $c = 1$ (Bjorken 1964). This convention makes the equations look a great deal simpler and we will adopt it in this section and in Appendix 2.

For a self-interacting real scalar inflaton field φ the standard form of the Lagrangian and action are

$$L = \frac{1}{2} \varphi_{,\mu} \varphi_{,\nu} g^{\mu\nu} - V(\varphi), \quad S = \int L \sqrt{-|g|} d^4x. \quad (19.9)$$

Note that in this chapter only we use a different notation convention for the metric determinant than we did previously in Chaps. 4 and 6; there we took $|g|$ to be the absolute value of the determinant: here we keep the sign of $|g|$ explicit. We do this to clarify some algebraic manipulations involving signs, especially in Appendix 2. Here V is a self-interaction potential function of the field, to be specified; for example the choice of a quadratic potential describes a massive scalar field (Bjorken 1965). The equation of motion for φ is then found using the standard Euler–Lagrange procedure

(Appendix 2) and the Lagrangian in (19.9) to be

$$\frac{1}{\sqrt{-|g|}} \left(\sqrt{-|g|} \varphi_{,\mu} g^{\mu\nu} \right)_{,\nu} + \frac{\partial V}{\partial \varphi} = 0. \quad (19.10)$$

The first term we recognize as the covariant Laplacian, discussed in Chap. 6. We assume the inflaton field is at least approximately uniform, as appropriate to an isotropic and homogeneous universe, and use the FLRW metric to obtain the dynamical equation

$$\ddot{\varphi} + 3H\dot{\varphi} + \frac{\partial V}{\partial \varphi} = 0, \quad H \equiv \frac{\dot{a}}{a}. \quad (19.11)$$

Here H is the usual Hubble function and the dot denotes a time derivative. This equation is the same as that of a particle acted on by a driving force proportional to the derivative of the potential function and a damping term proportional to the Hubble function. See Exercises 19.4 and 19.5.

The energy-momentum tensor for the scalar field is fundamentally important since it is the source of the gravitational field. It may be defined as the variational derivative of the Lagrangian with respect to the metric and found by standard methods to give the energy density and pressure of a uniform scalar field (Appendix 2). The result is

$$\rho_{\varphi} = \frac{1}{2}\dot{\varphi}^2 + V, \quad p_{\varphi} = \frac{1}{2}\dot{\varphi}^2 - V, \quad \frac{p_{\varphi}}{\rho_{\varphi}} = -\frac{V - \dot{\varphi}^2/2}{V + \dot{\varphi}^2/2}. \quad (19.12)$$

These relations make it clear that a scalar field can act like a constant vacuum density fluid if the field is spatially uniform and has negligible time variation compared to the potential V ; that is its effective equation of state is $p = -\rho$ and the energy density is dominantly due to the self-interaction, $\rho \cong V$.

The general scenario usually assumed for the behavior of the inflaton field and the inflating universe is that the field begins at some initial value and subsequently decreases slowly, thereby acting like a constant density vacuum fluid. This behavior is analogous to a mass rolling down a hill and is naturally called “slow roll.” During

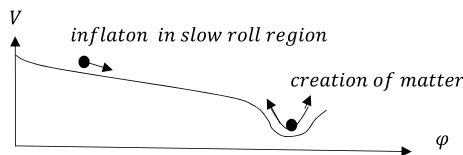


Fig. 19.2 Sketch of an example potential for inflation. The inflaton field “rolls” slowly down the potential hill, causing the universe to expand enormously, then ends up oscillating at the bottom of the potential well and gives rise to all matter while the large energy of the inflaton field itself almost completely vanishes. The transition is called reheating, a misnomer since there was likely no previous heating

the slow roll the universe expands enormously with an approximately exponential scale factor. Finally, the inflaton field decreases to a point where it oscillates at the bottom of a potential well of the potential V , and its energy is somehow converted into the particles of matter that now exist in the universe, that is the quarks and leptons and other constituents of the standard model (Quigg 2006; Linde 2007). To build a specific theory one must choose a potential that allows this sort of behavior; there are an infinite number of potential choices possible since there is little or no direct observational data to constrain the choice. Figure 19.2 is a generic sketch (Kinney 2002).

Some specific examples of inflaton potentials are

$$V = \frac{1}{2}m^2\varphi^2 \quad \text{mass term} \quad (19.13a)$$

$$V = \lambda\varphi^4 \quad \text{simple ad hoc} \quad (19.13b)$$

$$V = \lambda(\varphi^2 - M^2)^2 \quad \text{Higgs potential} \quad (19.13c)$$

See Exercise 19.8. These potentials and many others have been studied by theorists with the goal of giving predictions that may be compared with observation (Liddle 1999; Linde 2007; Kinney 2002). Almost needless to say, no inflaton field has been detected in the laboratory.

19.3 Origin of Structure

In the previous section we focused on the way that an inflaton field can act like a constant vacuum density and cause inflation, thus resolving the horizon problem. There is another important feature of the inflation scenario: it provides a mechanism for the origin of the structure seen in the later universe, including the anisotropy spectrum of the CMB that we discussed in Chap. 17.

As in the preceding section we assume that the inflaton field is nearly uniform so that its energy density is also nearly uniform and it acts much like a constant vacuum density, that is a cosmological constant. However quantum theory does not allow a strictly uniform field since that is not consistent with the uncertainly principle. Thus we must consider quantum fluctuations in the inflaton field. It is a remarkable feature of the inflaton scenario that fluctuations must occur, and fluctuations of very small size and magnitude can grow as the universe expands, and these become the seeds for all the structure we observe in the universe. See Exercise 19.9. Specifically, the fluctuations produce a spatial variation in energy density, and that gives rise to anisotropies in the CMB, and subsequently produces concentrations of energy that are seeds for the formation of stars and galaxies and clusters of galaxies. In particular the anisotropies of the CMB can be understood and compared with the observed spectrum.

Our goal in this section is quite modest since the theory of structure formation via inflation is large in scope (Kinney 2002; Linde 2007). Our aim is only to obtain a rough qualitative understanding of the behavior of the fluctuations as the universe expands. To do this we consider two length scales: one scale is the Hubble length $L_H = c/H$, which determines a fundamental causal region, and the other is the physical spatial size of the fluctuations.

A basic property of the FLRW geometry is that the physical separation L of co-moving objects increases proportional to the scale factor, and is given by

$$L = a\sigma, \quad \sigma = \text{constant coordinate separation.} \quad (19.14)$$

In this section we will take the scale factor to be 1 at some arbitrary initial time, rather than at the present time t_0 , and we will *not* use units with $c = 1$. Thus the velocity of separation between co-moving points and a Hubble type law are

$$v = L' = a'\sigma = \frac{a'}{a}(a\sigma) = HL = \frac{c}{L_H}L, \quad \text{prime denotes } \frac{d}{dt}. \quad (19.15)$$

From this it is clear that the Hubble length defines a causal region: within this region co-moving objects move apart at less than c and outside of it they move apart faster than c and can have no causal influence on each other. Thus the Hubble length defines a cosmological horizon. (Be aware that the word horizon is used for a number of different things in physics and cosmology.) During the inflationary era the Hubble length is constant at $L_H = c/H_I$ so the causal region does not change. During later times such as the radiation and matter dominated eras it grows linearly with time, specifically

$$L_H = 2ct, \quad \text{radiation era, } L_H = \left(\frac{3}{2}\right)ct, \quad \text{matter era.} \quad (19.16)$$

Loosely speaking nothing larger than the Hubble distance can be considered coherent.

The second length to consider is the spatial size of the inflaton field fluctuations. We may think of a fluctuation as composed of modes with wavelengths λ . It is clear from the above comments that only modes with λ less than about the Hubble length can have a direct physical effect in producing structure; those with larger λ act like constants and have negligible gradients, so do not produce physical effects such as concentrations of energy density. This is indicated in Fig. 19.3. As the universe expands the physical wavelength λ of a mode will stretch like a wavelength of light for the same reasons we discussed in Sect. 13.3; that is, the mode wavelength will increase proportional to the scale factor, which increases very rapidly during inflation.

Thus an initial mode wavelength λ_i will grow according to

$$\lambda(t) = a(t)\lambda_i, \quad \lambda_i \text{ initial wavelength.} \quad (19.17)$$

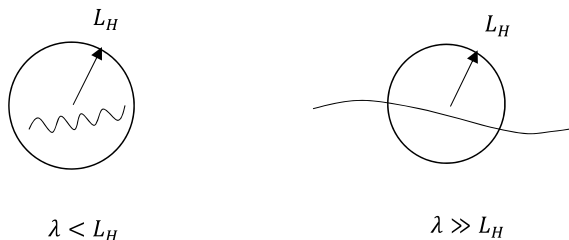


Fig. 19.3 Fluctuation on the left with a wavelength less than the Hubble length (circle) produces interesting physical effects, while the fluctuation on the right with wavelength much greater than the Hubble distance (circle) does not

The wavelength will thus quickly become larger than the constant Hubble length and will then not produce physical effects such as density variations since the change within the Hubble length is small. This is often referred to as “freezing of the mode as it crosses the horizon”—that is as λ expands outside the Hubble length. Figure 19.4 shows a wavelength beginning at less than the Hubble length and rapidly growing to exceed it.

It is also important to compare the wavelength and Hubble length in the radiation era. In the radiation era the scale factor is proportional to the square root of the time as in (17.12), so the Hubble length increases linearly with time and is proportional to the square of the scale factor. At some time the Hubble length must therefore become larger than the wavelength of the mode, which only increases linearly with the scale factor; we say the mode then “reenters the horizon” or is no longer frozen. This situation is illustrated in Fig. 19.4.

After reentering the causal regions the fluctuation mode can interact with material and geometry in the universe to act as a seed for future structure.

It is possible to illustrate this process more elegantly with a logarithmic plot that also includes the matter era. Here are the relevant relations for the scale factor and the Hubble length using the usual rough approximations for the scale factor,

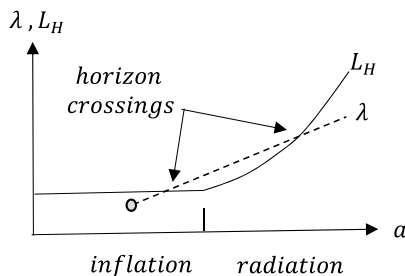


Fig. 19.4 Hubble length and the wavelength of a mode versus the scale factor. The wavelength increases during inflation to exceed the Hubble length. Later the Hubble length increases faster than the wavelength and the mode reenters the causal region

$$\text{Inflation: } a = Ae^{H_I t}, \quad L_H = \frac{c}{H_I}, \quad \ln L_H = \ln\left(\frac{c}{H_I}\right) \quad (19.18a)$$

$$\text{Radiation: } a = Bt^{1/2}, \quad L_H = \frac{2c}{B^2}a^2, \quad \ln L_H = 2 \ln a + \ln\left(\frac{2c}{B^2}\right) \quad (19.18b)$$

$$\text{Matter: } a = Ct^{2/3}, \quad L_H = \frac{3c}{2C^{3/2}}a^{3/2}, \quad \ln L_H = \frac{3}{2} \ln a + \ln\left(\frac{3c}{2C^{3/2}}\right) \quad (19.18c)$$

$$\text{De Sitter: } a = De^{H_d t}, \quad L_H = \frac{c}{H_d}, \quad \ln L_H = \ln\left(\frac{c}{H_d}\right). \quad (19.18d)$$

Notice that the logarithmic relations in (19.18) are all conveniently linear.

Figure 19.5 is similar to Fig. 19.4 and shows a plot of the Hubble length from (19.18) as four straight lines. It is clear from the figure how the wavelength of a mode leaves the Hubble region during inflation and reenters during the radiation or matter era. The part of the plot where the wavelength is below the Hubble curve is where the fluctuations interact with the material of the universe and have an effect on density variations. Note that the longer wavelengths reenter and become effective at later times than the shorter wavelengths.

We have only considered in this brief discussion some aspects of the fluctuations leading to structure. Calculation of the nature and magnitude of the fluctuations and their effect on the CMB and later structure is beyond our present scope. For a comprehensive discussion the reader may consult Linde (2007) and Kinney (2002).

Finally we note that some theorists speculate that the same fluctuation mechanism that produces structure should also give rise to many entire universes that are not connected in any evident way to our own (Linde 2007). This idea faces the serious objection that such other universes may not be subject to observational test and thus not subject to scientific study. In the next section we will mention one aspect of the multiverse idea, that it is related to the calculability of the values of some of the fundamental constants of nature.

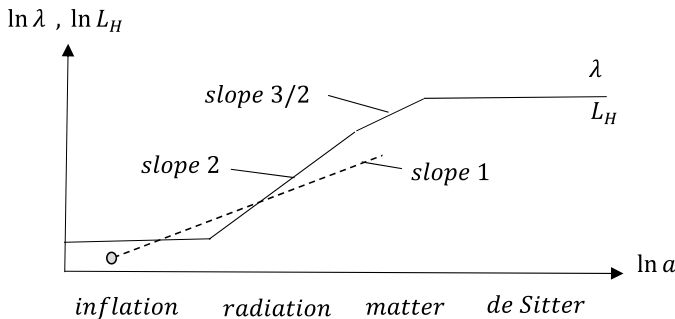


Fig. 19.5 This shows Fig. 19.4 in logarithmic form and includes the matter era. The Hubble length and the mode wavelength are plotted versus the scale factor. The causal region is where $\lambda < L_H$. The figure is a qualitative sketch and very much not to scale

19.4 The Physical Nature of Dark Energy

In the preceding chapters we have taken the dark energy to be synonymous with vacuum energy or the cosmological constant λ , whose equivalent energy density is

$$\rho_V = \frac{c^4}{8\pi G} \Lambda. \quad (19.19)$$

This is completely consistent with observations. In particular observations before about 2020 placed the w parameter for dark energy in the interval $-0.9 > w > -1.1$. A time variation of w has been searched for and not observed. In the context of general relativity theory dark energy may be considered to be the constant energy of the vacuum, that is space which is empty of all other matter or energy.

In quantum field theory the vacuum also has a constant nonzero energy density, which we mentioned briefly in Chap. 14. But the vacuum of quantum field theory has an infinite density, which of course is not to be taken seriously. As we mentioned before, attempts to understand this divergent theoretical energy using simple dimensional cutoff arguments give a finite estimate for the vacuum energy, specifically the Planck energy divided by the Planck distance cubed (see Sect. 19.6). But this value disagrees with the observed value by more than 120 orders of magnitude, which some consider a theoretical catastrophe (Adler 1995). If supersymmetry is invoked in the field theory the disagreement is less, but still about 60 orders of magnitude, and remains nonsense. Thus the quantum field vacuum does not appear to be understood and does not appear to be related to the general relativity vacuum in any evident way. It is reasonable for us to ignore it when working in cosmology.

It is also possible to interpret the dark energy as “stuff” whose energy density is very nearly constant. That is, the present dark energy could be an analog of the inflaton field used in inflation theory as we discussed in the preceding section, but with a vastly smaller energy density. Some versions of the stuff are referred to as quintessence, and there are many other speculative ideas and names. In addition there are speculations on a modification of gravitational theory, and many papers and books have been written on the subject (Amendola 2010).

If we accept the cosmological constant as the explanation for dark energy then its qualitative nature is not at all mysterious since the Einstein equations are completely consistent with a cosmological constant; indeed the equations almost demand it! However the very small numerical value of the cosmological constant remains an interesting and perhaps deep question. The value is very important since it is related to the size and age of the actual universe: the value of the Hubble constant in the Λ CDM universe is roughly the asymptotic or de Sitter value $\sqrt{3/\Lambda}$.

There are essentially two viewpoints one can take regarding the numerical value of the cosmological constant: one may try to calculate it from some fundamental theory or principle, or one may simply take it as a fact of nature, an accidental number. The first of these has not been successful to date. The second has led to interesting questions and may be related to the idea of a multiverse.

History provides an analogy. In the sixteenth century Johannes Kepler attempted to calculate the relative sizes of the orbits of the planets according to ideas popular at the time, that is in terms of Euclidean geometry and Platonic solids. It only became clear many years later that his attempt was doomed since we now know there are many stellar planetary systems in which the relative sizes of the planetary orbits are different than in our solar system; the orbital ratios are accidents of initial conditions and complex dynamics. Some theorists have used such historical facts in support of the multiverse idea, that there are many universes in which the cosmological constant can have diverse values; the value is an accident of initial conditions (Susskind 2005).

We have focused on the cosmological constant numerical value in the above paragraphs, but it is clear that the idea of a multiverse is also relevant to the values of the other fundamental constants in our universe, in particular the parameters of the standard model of particle physics, such as particle charges and masses and mixing parameters: these might all be accidents and not calculable.

The idea of a multiverse has produced much controversy and objections, the deepest of which is the question of whether it can make testable predictions about the one universe we actually observe, and thus whether it is in any way relevant to science. There is now a large literature and diverse opinions on the subject (Vaas 2010).

19.5 The Physical Nature of Dark Matter

At present all the relevant observations of dark matter concern its gravitational effects. The physical nature of dark matter is unknown. The theory and observational search for dark matter are one of the most active research areas in physics. Of the many theoretical guesses as to its nature we will only mention a few of the most popular or interesting (Randall 2018).

Large macroscopic bodies such as burnt out stars and stellar mass black holes were some of the first things considered as dark matter candidates. They have been searched for and not found. There is a further problem with any dark matter candidate composed of ordinary baryons, that nucleosynthesis theory and observation suggest that there cannot be large numbers of such objects (Misner 1973; Drees 2018).

Particles of various kinds that have not yet been seen in the lab are natural candidates for dark matter. One of the most popular types is the so-called weakly interacting massive particle, the WIMP; such particles are particularly attractive to some theorists since they are a natural part of supersymmetric (SUSY) quantum field theory. However for decades there have been many active searches for WIMPs in the laboratory, all with negative results. Moreover there is as yet no experimental evidence for SUSY. The possibility remains that WIMPs with large enough mass to have eluded detection could be the dark matter, and the search continues (Drees 2018).

For some general relativity theorists there is a dark matter candidate that is particularly interesting. As we discussed previously black holes of sufficiently small mass should emit Hawking radiation and grow yet smaller and thus radiate faster. It can

be argued that the end result could be a black hole remnant of about the Planck mass, that is of order 10^{19} GeV (Adler 2001). Such a particle would have only a gravitational interaction and also a very low number density. As a result it would be extraordinarily difficult to detect in the lab or by any means other than large scale gravitational effects; it would be an experimentalist's nightmare. Moreover Hawking radiation has been searched for and not yet seen, so black hole remnants are quite speculative. We will discuss them further in the next section on quantum effects and in Appendix 3.

An entirely different possibility is that dark matter simply does not exist, that general relativity is not the correct or complete theory of gravity and the many observations that purport to measure dark matter on a galactic scale and larger are not being interpreted correctly. One such type of theory is called MOND for “modified Newtonian dynamics” (Milgrom 2014). Another is called conformal gravity (Mannheim 2011). Such theories are not now in the mainstream so we refer the reader to the above references.

19.6 The Planck Era and Quantum Physics

Max Planck discovered the quantum constant $\hbar$ when studying black body radiation. He realized that the constants $\hbar$ and c and G determine a natural scale, now called the Planck scale (Planck 1899). The values of the constants are

$$\begin{aligned} c &= 3.00 \times 10^8 \text{ m/s}, \quad \hbar = 1.05 \times 10^{-34} \text{ J s}, \\ G &= 6.67 \times 10^{-11} \text{ N m}^2/\text{kg}^2. \end{aligned} \quad (19.20)$$

They lead to the Planck length, time, mass and energy values

$$\begin{aligned} L_P &= \sqrt{\frac{G\hbar}{c^3}} = 1.6 \times 10^{-35} \text{ m}, \quad T_P = \frac{L_P}{c} = \sqrt{\frac{G\hbar}{c^5}} = 0.54 \times 10^{-43} \text{ s}, \\ M_P &= \frac{\hbar}{L_P c} = \sqrt{\frac{\hbar c}{G}} = 2.2 \times 10^{-8} \text{ kg}, \\ E_P &= M_P c^2 = \sqrt{\frac{\hbar c^5}{G}} = 2.0 \times 10^9 \text{ J} = 1.2 \times 10^{19} \text{ GeV}. \end{aligned} \quad (19.21)$$

From the way it is constructed the Planck scale should be relevant when the system considered is quantum mechanical ($\hbar$), involves high velocities and large energies (c), and in which gravity is strong (G). One such system is the very early universe. Another is the evaporation of a black hole. The collision of particles in a laboratory at the Planck energy would be very interesting but unfortunately would require an

accelerator larger than the solar system. Even the highest energy cosmic rays ever observed have much less than Planck energy.

It is believed by many cosmologists that there was an initial Planck era before inflation in which quantum effects were important and the universe was governed by Planck scale phenomena, that is quantum gravity. Much work has gone into constructing a quantum theory of gravity appropriate to the Planck scale, but with little or no predictive success (Rovelli 2008; Frignanni 2011). For example the strings of superstring theory are of Planck size. We cannot go into theories such as string theory or loop quantum gravity here, but will instead content ourselves with the more modest task of obtaining a generalized uncertainty principle and using it to show that the Planck length arises naturally as a minimum meaningful distance when we combine the ideas of quantum mechanics and basic ideas of gravity and general relativity.

We will first generalize the uncertainty principle (UP) of quantum mechanics to include gravitational effects and obtain a generalized uncertainty principle (GUP). Our argument will be rough order of magnitude and largely based on the Heisenberg uncertainty principle of quantum theory, which we will now recall. General principles of optics and quantum mechanics tell us that if we measure the position of a particle, such as an electron, with a photon of wavelength λ we cannot expect better precision than about λ , which we express as

$$\Delta x_H \gtrsim \lambda. \quad (19.22)$$

A photon of wavelength λ has a momentum of $p = 2\pi\hbar/\lambda$, and when it interacts and scatters from the particle a significant fraction of this momentum will generally be given to the particle $\Delta p \approx p = \hbar/\lambda$. This makes its momentum uncertain to roughly

$$\Delta p \approx \hbar/\lambda. \quad (19.23)$$

Combining these last two equations we obtain

$$\Delta x_H \Delta p \gtrsim \hbar, \quad \text{or} \quad \Delta x_H \gtrsim \hbar/\Delta p, \quad (19.24)$$

which is the well-known Heisenberg uncertainty principle. Figure 19.6 shows a picture of the process of measuring the particle position.

But this illustration of the uncertainty principle ignores gravity. The particle will also interact gravitationally with the photon which produces spacetime curvature, and this should produce an additional uncertainty in the position of the particle. If the wavelength of the photon is small and its momentum and energy are large this interaction can become too large to ignore. We can estimate the effect by a heuristic dimensional argument. Let us include the gravitational effect and call the extra term Δx_g . This gravitational term should obviously be proportional to the gravitational

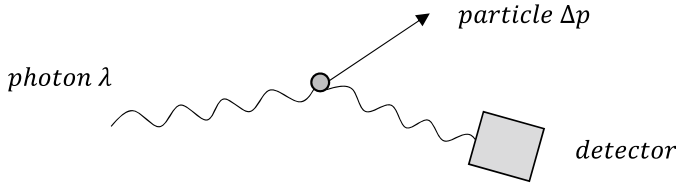


Fig. 19.6 The Heisenberg microscope uses a photon to measure the position of a particle with inescapable imprecision or uncertainty due to the wave nature of light

constant G . It should also be proportional to the energy E of the photon since energy is the source of gravity; this implies that it should be proportional to the momentum $p = E/c$ of the photon which we take to be comparable to the momentum transfer $\Delta p \approx p$. We thus have

$$\Delta x_g \propto G \Delta p. \quad (19.25)$$

In order to give the gravitational term the correct dimensions we note that G is proportional to the square of the Planck length and that a momentum over $\hbar$ is a distance. From this we see that we may rewrite (19.25) in dimensionally correct form as

$$\Delta x_g \sim L_P^2 \frac{\Delta p}{\hbar}, \quad L_P^2 = \frac{G \hbar}{c^3}. \quad (19.26)$$

Since this is only a heuristic rough estimate we should think of L_P as a small length of order the Planck length.

Adding the additional gravitational uncertainty (19.26) to the Heisenberg uncertainty (19.24) we have the GUP

$$\Delta x_{\text{tot}} \gtrsim \frac{\hbar}{\Delta p} + L_P^2 \frac{\Delta p}{\hbar}, \quad \Delta x_{\text{tot}} \Delta p \gtrsim \hbar \left[1 + \left(\frac{\Delta p}{\hbar} \right)^2 L_P^2 \right] \text{ GUP}. \quad (19.27)$$

It is obvious that the extra gravitational term in (19.31) and (19.32) is utterly unimportant at present laboratory energies since the Planck length is so small.

The GUP has a remarkable consequence for the nature of spacetime. If the photon momentum is very small then the particle position is imprecise because the long photon wavelength gives poor resolution. If the photon momentum is chosen very large then its gravitational field makes the particle position very imprecise. Between the two extremes there is a minimum position uncertainty, as shown in Fig. 19.7.

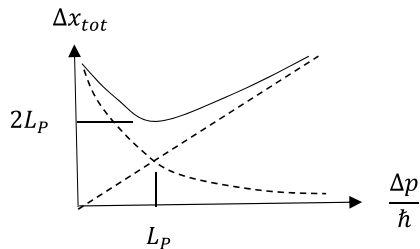


Fig. 19.7 The minimum uncertainty is of order the Planck length

From (19.27) we find the minimum to be

$$\Delta x_{\text{tot}} \approx 2L_P \quad \text{for} \quad \frac{\Delta p}{\hbar} \approx L_P. \quad (19.28)$$

This means that we cannot localize the position of a particle to better than about the Planck length, and may do that by using photons with about the Planck energy. The Planck length thus appears as a minimum distance that has physical meaning. In this sense space has a granular structure. Consequently we also expect that the Planck time is the minimum time that has physical meaning, so the history of the universe may only go back to about the Planck time.

The GUP was first obtained in studies in string theory, but it was soon realized that it should be understandable on more general and basic grounds; there are indeed a large number of ways to obtain it on somewhat more convincing grounds than the heuristics we have used here (Scardigli 1999; Adler 1999). Other analyses using the path integral approach to quantum theory lead to analogous conclusions, namely that spacetime at small distances and times undergoes quantum fluctuations, and at the Planck scale the fluctuations are of the same order as the distances involved; spacetime becomes a sort of foam (Misner 1973). Thus classical spacetime has no meaning at this scale, and must be replaced by something more fundamental, such as a spacetime amplitude or wave function.

In Appendix 3 we will discuss the possibility that the GUP might have observable consequences involving evaporating black holes and dark matter.

Appendix 1: Power Law Inflation

Power law inflation is probably the simplest model of inflation and a good pedagogical example (Peebles 1993). Moreover it provides an intrinsically interesting example of how one can choose a scale factor and derive from it an effective equation of state for a corresponding fluid. In this appendix we will explicitly obtain the parameter w in the equation of state $p = w\rho$ for power law inflation.

The basic assumption is that the scale factor is equal to some power of time with a rather large exponent $m \gg 1$,

$$a = Ct^m, \quad C = \text{const.} \quad (19.29)$$

Then the horizon problem is trivially resolved since the integral in the horizon condition (19.1) diverges, assuming of course that the initial inflation time t_1 is very small or zero.

To see what sort of fluid this scale factor might correspond to we refer to Sect. 14.4, in particular to (14.15b) which gives the behavior of the energy density of a fluid as a function of the scale factor as the universe expands. We rewrite (14.15b) as

$$\rho = \frac{D}{a^{3(1+w)}}, \quad D = \text{const.} \quad (19.30)$$

It is clear that for such a fluid the appropriate Friedmann equation is (14.17) with no cosmological constant or curvature; we express it as

$$\frac{a'^2}{a^2} = \frac{E}{a^{3(1+w)}}, \quad E = \text{const.} \quad (19.31)$$

Finally, to relate the w of the fluid to the power m we substitute the scale factor (19.29) into (19.31) and have

$$\left(\frac{m}{t}\right)^2 = \frac{E}{(At^m)^{3(1+w)}}. \quad (19.32)$$

Equating the powers of t we get the simple relation

$$w = -1 + \frac{2}{3m}. \quad (19.33)$$

Thus the fluid that produces power law inflation has an equation of state w parameter that is a little larger than -1 . In this sense it behaves similarly to cosmological constant vacuum energy, which has $w = -1$.

Appendix 2: Scalar Field Theory

In this appendix we will simplify notation by setting the constants c and $\hbar$ equal to one, as is commonly done in particle physics and inflation theory. With this choice every quantity in the theory can be chosen to have the dimension of distance to some power.

Scalar field theory is generally based on a scalar Lagrangian (Bjorken 1965). From a scalar field φ we may form two types of scalars, one composed of first derivatives, and the other some scalar function of the field. Thus we take the Lagrangian to be the scalar

$$L = \frac{1}{2} \varphi_{,\mu} \varphi_{,\nu} g^{\mu\nu} - V(\varphi). \quad (19.34)$$

Here the potential function V is a self-interaction to be determined. For example, for the special case of a free particle of mass m the appropriate potential is

$$V = \frac{1}{2} m^2 \varphi^2, \text{ free particle of mass } m. \quad (19.35)$$

To obtain the equations of motion for the field φ we define the action S to be the integral of the Lagrangian density $\mathcal{L}$ over all spacetime,

$$S = \int \mathcal{L} d^4x, \quad \mathcal{L}(\varphi, \varphi_{,\mu}) = L(\varphi, \varphi_{,\mu}) \sqrt{-|g|}. \quad (19.36)$$

As noted in the text we keep the sign of the metric determinant $|g|$, which is negative with our metric choice, explicit in this appendix in order to make the following sign manipulations clear. The equations of motion are obtained by setting the variational derivative of this action with respect to the field equal to zero; that is, the action is to be extremized. The variation of the action with respect to a change $\delta\varphi$ in the field is computed in the standard way as

$$\begin{aligned} \delta S &= \int \left[\left(\frac{\partial \mathcal{L}}{\partial \varphi} \right) \delta\varphi + \left(\frac{\partial \mathcal{L}}{\partial \varphi_{,\mu}} \right) \delta\varphi_{,\mu} \right] \delta\varphi d^4x \\ &= \int \left[\left(\frac{\partial \mathcal{L}}{\partial \varphi} \right) \delta\varphi + \frac{\partial}{\partial x^\mu} \left(\frac{\partial \mathcal{L}}{\partial \varphi_{,\mu}} \delta\varphi \right) - \left(\frac{\partial}{\partial x^\mu} \frac{\partial \mathcal{L}}{\partial \varphi_{,\mu}} \right) \delta\varphi \right] d^4x = 0, \\ &= \int \left[\left(\frac{\partial \mathcal{L}}{\partial \varphi} \right) - \left(\frac{\partial}{\partial x^\mu} \frac{\partial \mathcal{L}}{\partial \varphi_{,\mu}} \right) \right] \delta\varphi d^4x = 0. \end{aligned} \quad (19.37)$$

In this we have integrated by parts in the second line; in the third line the middle term has been discarded since it leads by Gauss's Theorem to a surface integral, which is zero if the volume is taken as all of spacetime. The square bracket in the integral in the last line must therefore be zero; this gives the canonical Euler-Lagrange equations in covariant form,

$$\left(\frac{\partial \mathcal{L}}{\partial \varphi} \right) - \frac{\partial}{\partial x^\mu} \left(\frac{\partial \mathcal{L}}{\partial \varphi_{,\mu}} \right) = 0. \quad (19.38)$$

For the scalar field Lagrangian in (19.34) this may be written as

$$\frac{1}{\sqrt{-|g|}} \left(\sqrt{-|g|} \varphi_{,\mu} g^{\mu\nu} \right)_{,\nu} + \frac{\partial V}{\partial \varphi} = 0. \quad (19.39)$$

The first term is the covariant Laplacian we discussed in Chap. 6.

For the present case the metric is taken to be that of de Sitter space in Cartesian coordinates as discussed in Sects. 16.7 and 19.1. Then $\sqrt{-|g|} = a^3$ and we find from (19.34)

$$\varphi_{,0,0} - \frac{1}{a^2} \varphi_{,i,i} + 3 \frac{a_{,0}}{a} \varphi_{,0} + \frac{\partial V}{\partial \varphi} = 0. \quad (19.40)$$

This gives (19.11) in the text if the field is assumed to be uniform in space.

The energy momentum tensor for the scalar field may be obtained in a similar way. It is conveniently defined in terms of the derivative of the field action with respect to the metric. This definition is motivated by the fact that the Einstein tensor is the variational derivative with respect to the metric of the gravitational action (Adler 1975). We thus examine the quantity

$$\begin{aligned} \frac{\partial}{\partial g^{\mu\nu}} \left(L \sqrt{-|g|} \right) &= \sqrt{-|g|} \left(\frac{\partial L}{\partial g^{\mu\nu}} \right) + L \left(\frac{\partial \sqrt{-|g|}}{\partial g^{\mu\nu}} \right) \\ &= \sqrt{-|g|} \left(\frac{\partial L}{\partial g^{\mu\nu}} \right) + \frac{L}{2\sqrt{-|g|}} \left(\frac{\partial (-|g|)}{\partial g^{\mu\nu}} \right), \end{aligned} \quad (19.41)$$

and will identify the energy momentum tensor from it.

The derivative that appears in the last term in (19.41) is slightly tricky to evaluate. Recall that in Chap. 6 we dealt with a similar derivative—that of the metric determinant with respect to the metric tensor, whereas now we want the derivative with respect to the inverse metric tensor. Let us then consider the determinant of the inverse metric tensor and call it $|\hat{g}|$; since the determinant of the inverse of a matrix is equal to the inverse of the determinant we have $|\hat{g}| = 1/|g|$. It is a well-known property of matrices that the determinant of a matrix can be expressed in terms of its cofactor, and the inverse of the matrix can also be expressed in terms of the cofactor, as follows

$$\hat{g} = g^{\mu\nu_0} C_{\mu\nu_0}, \quad g_{\mu\nu} = \frac{C_{\mu\nu_0}}{|\hat{g}|}, \quad C_{\mu\nu} = \text{cofactor } g^{\mu\nu}. \quad (19.42)$$

(See Chap. 6, and note that ν_0 is a fixed index and is not to be summed over.) From these two equations we obtain the derivative of the determinant $\hat{g}$ with respect to $g^{\mu\nu}$ as

$$\frac{\partial |\hat{g}|}{\partial g^{\mu\nu}} = C_{\mu\nu} = g_{\mu\nu} |\hat{g}|. \quad (19.43)$$

But we know that $|\hat{g}| = 1/|g|$, so by substituting this in the last equation we finally obtain the desired derivative

$$\frac{\partial}{\partial g^{\mu\nu}} \frac{1}{|g|} = g_{\mu\nu} \frac{1}{|g|}, \quad \frac{\partial |g|}{\partial g^{\mu\nu}} = -|g| g_{\mu\nu}. \quad (19.44)$$

Substituting this into (19.41) we obtain

$$\frac{\partial}{\partial g^{\mu\nu}} (L\sqrt{-|g|}) = \sqrt{-|g|} \left[\left(\frac{\partial L}{\partial g^{\mu\nu}} \right) - \frac{L}{2} g_{\mu\nu} \right]. \quad (19.45)$$

Accordingly we take the square bracket in (19.45) to be the energy momentum tensor $T_{\mu\nu}$ up to a constant factor. The lower index and the mixed index energy momentum tensor are thus

$$T_{\mu\nu} = C \left[\left(\frac{\partial L}{\partial g^{\mu\nu}} \right) - \frac{L}{2} g_{\mu\nu} \right], \quad T^{\mu}_{\nu} = C \left[\frac{\partial L}{\partial g^{\alpha\nu}} g^{\alpha\mu} - \frac{L}{2} g^{\mu}_{\nu} \right], \quad (19.46)$$

where the constant C is to be determined.

For a uniform scalar field the energy density T^0_0 and the pressure $-T^i_i$ for a uniform field are then, from (19.34) and (19.46),

$$\rho_{\varphi} = \frac{1}{2} \dot{\varphi}^2 + V, \quad p_{\varphi} = \frac{1}{2} \dot{\varphi}^2 - V, \quad (19.47)$$

where we have chosen $C = 2$ to make $\rho = V$ for the case of a uniform static field. Equation (19.12) in the text is thus verified. As noted there we see that if the time variation of the scalar field is small then the w parameter in the effective equation of state is near -1 and the uniform scalar field acts like a constant vacuum energy density.

Appendix 3: Black Hole Remnants as Dark Matter

Recall that in Chap. 10 we discussed in a heuristic way how black holes are theorized to have a nonzero temperature, the Hawking temperature, and thus radiate like black bodies. There is no conserved quantum number associated with a black hole, so one might expect that it should radiate away completely, leaving behind only the radiated particles. However there is a plausible argument one can make that a remnant should

be left behind. The GUP may prevent total evaporation in exactly the same way that the uncertainty principle prevents a hydrogen atom from total collapse: the complete decay of a black hole is prevented, not by symmetry, but by dynamics, as a minimum size and mass are approached (Adler 2001). See Exercise 19.11.

We may use the GUP to derive a modified black hole temperature exactly as we derived the Hawking temperature in Chap. 10. The basic idea is the same; a virtual pair of charged particles and a photon are formed near the black hole surface, the particles are absorbed by the black hole, and the photon is emitted as black body thermal radiation as shown in Fig. 10.8. From (19.27) we solve for the emitted photon momentum in terms of the distance uncertainty, which we take to be the Schwarzschild radius $\Delta x = 2GM/c^2$, and obtain

$$p = \frac{\hbar \Delta x}{2L_P^2} \left[1 \pm \sqrt{1 - 4 \frac{L_P^2}{\Delta x^2}} \right] = Mc \left[1 - \sqrt{1 - \frac{M_P^2}{M^2}} \right]. \quad (19.48)$$

We have chosen the negative sign to agree with the results of Sect. 10.7. The energy of the photon is of course $E = pc$. Thus we estimate the temperature of the black hole to be $kT = E$ with

$$kT \approx E = Mc^2 \left[1 - \sqrt{1 - \frac{M_P^2}{M^2}} \right]. \quad (19.49)$$

It is easy to check that this agrees with the previous results of Chap. 10 for masses large compared to the Planck mass: we expand (19.49) and find

$$kT \approx \frac{M_P^2 c^2}{2M} = \frac{\hbar c^3}{2GM}, \quad (19.50)$$

which is roughly the same as our estimate (10.33) and the Hawking temperature (10.34). Notice also that the temperature (19.49) is well-behaved as the mass of the black hole approaches the Planck mass, whereas in the standard Hawking result it is infinite.

With the modified temperature (19.49) it is straightforward to calculate the entropy of a black hole in terms of its mass, and also its lifetime and the rate of energy radiated; all are well-behaved as the mass approaches the Planck mass and the black hole becomes a remnant. Figure 19.8 shows the mass as a function of time and compares the Hawking result to what we obtained with the GUP—corrected to agree with Hawking at $t = 0$.

In summary the picture that follows from the above calculation is that a small black hole, with temperature greater than the ambient temperature, should radiate photons, as well as other particles, until it approaches the Planck mass and size. At the Planck mass it ceases to radiate and its entropy reaches zero, even though

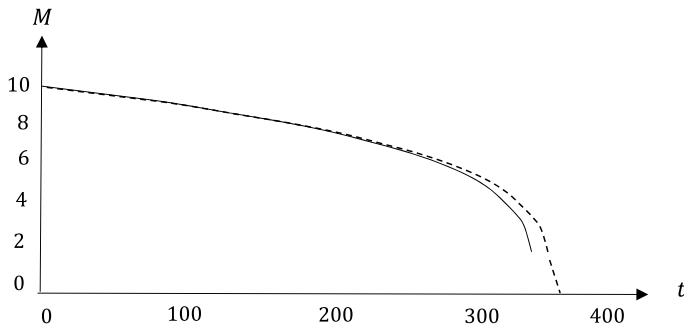


Fig. 19.8 The mass of a small black hole versus time. The mass is in units of the Planck mass and the time is in units of an arbitrary characteristic time. The upper dashed curve is the Hawking result and the lower is the result using the GUP

its temperature formally reaches the Planck energy! It then cannot radiate further and becomes an inert remnant, possessing only gravitational interactions. Note that the remnant need not have a classical black hole horizon structure. Such remnants may have been in existence since very early in the history of the universe and are a plausible dark matter candidate (Adler 2001).

As with most other calculations dealing with Hawking radiation we have not treated all of the gravitational aspects of the problem completely consistently. That is we have not taken account of the recoil of the black hole when radiating very high energy particles, possible quantization of the black hole mass and metric, and so forth. Thus, while we cannot expect our results to incorporate all aspects of quantum gravity near the Planck scale they do appear to be plausible and more consistent than the standard results.

The idea that dark matter is composed of black hole remnants may be attractive to theorists, but it is also a nightmare for experimentalists. Such remnants would interact very weakly, only via gravity. In particular the absorption cross section should be of order the Planck distance squared, very much less than that for WIMPS. Almost as bad, the number density would be quite low due to their large mass, which is huge by particle physics standards. The average number density of baryons in the universe is of order $1/\text{m}^3$. The requisite number density of black hole remnants should be of order $10^{-19}/\text{m}^3$ since the remnant mass is of order 10^{19} GeV. Even if the density is a million times larger near the center of galaxies this implies a number density of order $10^{-13}/\text{m}^3$. At that density the interparticle separation is of order 10^4 km. The chance of direct detection of such particles is clearly quite remote, leaving only observations of large scale gravitational effects.

Exercises

- 19.1 The equation of state parameter w can be estimated by observation, and is found to be close to -1 . Find in the literature how much it might differ from -1 according to current observations, then calculate from Appendix 1 what power law m parameter is consistent with this.

- 19.2 Use the scalar field Lagrangian in (19.9) to obtain the dynamical equations for the inflaton field in (19.10). This is also done in Appendix 2.
- 19.3 Work out the Lagrangian in (19.9) in ordinary units, that is in which $\hbar$ and c are not taken to be 1. Is it clear why natural units are preferred by theorists?
- 19.4 Consider one idealized over-simplified case for the inflaton field equation (19.11). Take the Hubble function to be zero and the potential V to be linearly decreasing, and solve for $\varphi(t)$. Does this clarify the physical role of V ?
- 19.5 Consider another idealized over-simplified case for (19.11). Take the Hubble function to be constant and the potential V to be zero, and solve for $\varphi(t)$. Does this clarify the physical role of H ?
- 19.6 The integral of the Lagrangian has the dimensions of an action in (19.9). Use this fact to work out the dimensions of the scalar field and the potential. Do this for both ordinary and natural units. Are they consistent with the energy density and pressure of the inflaton field in (19.12)?
- 19.7 Sketch the Inflaton Potentials in (19.13).
- 19.8 Take the inflaton self-interaction potential V to be that in (19.13a); show that the equation of motion is the Klein Gordon equation that describes a free particle of mass m in quantum field theory. You might want to assume the flat space of special relativity for simplicity.
- 19.9 Use dimensional analysis to show that the amplitude of fluctuations of the inflaton field during a hubble time period is of order $|\delta\varphi| \sim H$ (Linde 2007).
- 19.10 Consider a universe with a scale factor that is a power t^m as in (19.2) and Appendix 1. How would such a universe fit into the picture in Fig. 19.5 if m is large?
- 19.11 According to classical theoretical physics the hydrogen atom would not be stable since the electron would radiate energy and fall onto the proton. Use the UP to show how it is stable according to quantum theory, and obtain an estimate of the ground state energy.
- 19.12 What if the uncertainties in (19.27) add as squares? That is

$$\Delta x_{\text{tot}}^2 \gtrsim \left(\frac{\hbar}{\Delta p} \right)^2 + \left(L_P^2 \frac{\Delta p}{\hbar} \right)^2.$$

Which version do you think is a more reasonable guess? Show that the conclusions of Sect. 19.6 concerning a minimal length do not change significantly if this version is used.

References

- Abbott, B. P., et al. (2016). Observation of gravitational waves from a binary black hole merger. *Physical Review Letters*, 116, 061102.
- Abbott, B. P., et al. (2017). Observation of gravitational waves from a binary neutron star inspiral. *Physical Review Letters*, 119, 161101.
- Abbott, B. P., et al. (2019). Properties of the binary neutron star merger GW170817. *Physical Review X*, 9, 01101.
- Abbott, B. P. et al. (2017). *A standard siren measurement of the Hubble constant from GW170817*. arxiv.org/abs/1710.05835v1.
- Adler, R., Bazin, M. & Schiffer, M. (1975), *Introduction to general relativity* (2nd ed.). McGraw Hill.
- Adler, R. J., & Das, T. K. (1976). Charged black hole electrostatics. *Physical Review D*, 14, 2474.
- Adler, R. J. (1993). *Relativity, general theory*, in *McGraw Hill Encyclopedia of Physics* (2nd ed.). New York: McGraw Hill.
- Adler, R. J., Casey, B., & Jacob, O. (1995). Vacuum catastrophe: An elementary exposition of the cosmological constant problem. *American Journal of Physics*, 63(7), 620–626.
- Adler, R. J., & Santiago, D. (1999). On gravity and the uncertainty principle. *Physical Review A*, 14, 1371.
- Adler, R. J. (1999). Metric for an oblate earth. *General Relativity and Gravitation*, 31, 1999.
- Adler, R. J., & Silbergleit, A. S. (2000). A general treatment of orbiting gyroscopic precession. *International Journal of Theory Physics*, 39, 1287.
- Adler, R. J., Chen, P., & Santiago, D. (2001). The generalized uncertainty principle and black hole remnants. *General Relativity and Gravitation*, 33, 2101.
- Adler, R. J., & Overduin, J. (2005). The nearly flat universe. *General Relativity and Gravitation*, 37(9), 1491–1503.
- Adler, R. J., Bjorken, J. D., Chen, P. & Liu, J. S. (2005). Simple analytic models of gravitational collapse. *American Journal of Physics*, 73(12), January 2005.
- Adler, R. J. (2006). *Gravity*. In Gordon Fraser (Ed.), *The new physics for the twentieth-first century*. Cambridge University Press.
- Adler, R. J. (2006). Six easy roads to the planck scale. *American Journal Physics*, 78, 9.
- Amendola, L., & Tsujikawa, S. (2010). *Dark energy theory and observations*. Cambridge: Cambridge University Press.
- Arfken, G. (1970). *Mathematical methods for physicists* (2nd ed). Academic Press.
- Ashby, N. (2003). Relativity in the global positioning system. *Living Reviews in Relativity*, 6.
- Bekenstein, J. D. (1973). Black holes and entropy. *Physical Review D*, 7, 2333.

- Bennett, C., et al. (2003). First year wilkinson microwave anisotropy probe (MAP) observations. *Astrophysical Journal Supplement*, 148(1), 97–117.
- Bergmann, P. (1942). *Introduction to the theory of relativity*. Prentice Hall, 1942.
- Birrer, S. et al. (2020). *TDCOSMO IV: Hierarchical time-delay cosmography, joint inference of the hubble constant and galaxy density profiles*. Arxiv 2007.02941 v1, 6 Jul 2020.
- Bjorken, J. D. & Drell, S. D. (1964). *Relativistic quantum mechanics*. McGraw Hill.
- Bjorken, J. D. & Drell, S. D. (1965). *Relativistic quantum fields*. McGraw Hill.
- Boggess, N. W., et al. (1992). The COBE mission: Its design and performance two years after the launch. *Astrophysical Journal*, 397(2), 420.
- Bondi, H., & Gold, T. (1948). The steady state theory of the expanding universe. *MNRAS*, 108, 252.
- Burgay, M. (2012). The double pulsar system in its 8th anniversary. In: *Science with parkes at 50 years young*, 31 Oct.–4 Nov., (ATNF/CSIRO, Australia, 2012). [ADS].
- Carroll, B. W. & Ostlie, D. A. (2017). *An introduction to modern astrophysics* (2nd ed.). Cambridge University Press.
- Chalinoir, A. (2012). *CMB anisotropy science: A review*, Proc. IAU symposium No. 288.
- Chandrasekhar, S. (1939). *An introduction to the study of stellar structure*, (Dover 1939).
- Chen, G. C-F et al. (2019). A sharp view of HOLiCOW: H_0 from three time-delay gravitational lens systems with adaptive optics imaging, arxiv.org/abs/1907.02533.
- CMB Polarization. (2019). *CMB polarization*, cfa.harvard.edu.
- Cosmicweb.uchicago.edu. (2019). *From quantum foam to galaxies: formation of the large-scale structure of the universe*.
- Courant, R. & Hilbert, D. (1937). *Methods of mathematical physics*, Vol. I (Interscience publishers 1955).
- Crane, L. (2019). *Something is seriously wrong with our understanding of the cosmos*, New Scientist, July 11, 2019.
- Dar, A. (2006). The new astronomy. In G. Fraser (Ed.), *The new physics for the twentieth-first century*. Cambridge University Press.
- Dominguez, A. et al. (2019). *A new measurement of the Hubble constant and matter content of the universe using extragalactic background light gamma-ray attenuation*. arxiv.org/abs/1903.12097.
- Drees, M. (2018). *Dark matter theory*. arXiv:1811.06406v1, 2018.
- Dvorkin, C. et al. (2019). *Neutrino mass from cosmology: Probing physics beyond the standard model*. arxiv.org/abs/1903.03689.
- Eddington, A. S. (1923). *The mathematical theory of relativity*. Cambridge: Cambridge University Press.
- Eikenberry, S. et. al. (2019), *Astro 2020 Science White Paper, A Direct Measure of Cosmic Acceleration*. arxiv.org/abs/1904.00217.
- Einstein, A. & Lorentz, H. A. & Weyl H., & H. Minkowski, H. (1923), *The principle of relativity* (Dover, U.S. 1923). This contains various reprinted articles. In *Does the Inertia of a Body Depend Upon its Energy Content?* there is a wonderfully simple discussion of the famous equation $E = mc^2$.
- Einstein, A. (1934). *Essays in science*. U.S.: Philosophical Library.
- Eötvös, R. V., Pekár, V., & Fekete, E. (1922). Beiträge zum Gesetze der Proportionalität von Trägheit und Gravität. *Ann. Phys. (Leipzig)*, 68, 11–66.
- Event horizon telescope, EHT. (2019). First M87 event horizon telescope results. I. The Shadow of the Supermassive Black Hole. *The Astrophysical Journal* 87(1): L1.
- Everitt, F. et al. (2015). *Classical and quantum gravity* Vol. 32, 22 (IOP Bristol, 2015). The final report on the Gravity Probe B experiment to test gyroscope precession occupies the entire volume.
- Fischbach, M. et al. (2018). *A standard siren measurement of the Hubble constant from GW170817 without the electromagnetic counterpart*. arxiv.org/abs/1807.05667.
- Fixsen, D. J. et al. (1993). *The cosmic microwave background spectrum from the full COBE/FIRAS data set*. arXiv:astro-ph/9605054.
- Fraknoi, A. (2016). *Astronomy (The Textbook)* researchgate.net.

- Freedman, W. L. & Kolb, E. W. (2006). Cosmology. In G. Fraser (Ed.), *The new physics for the twentieth-first century*. Cambridge University Press.
- Freedman, W. et al. (2019). *The Carnegie-Chicago hubble program. VIII. An independent determination of the hubble constant base on the tip of the red giant branch*. arxiv.org/abs/1907.05922.
- Frignanni, V. R. (Ed.). (2011). *Classical and quantum gravity, theory, analysis and applications*. Nova Publishers.
- Godel, K. (1949). *An example of a new type of cosmological solutions of Einstein's field equations of gravitation*. *Reviews of Modern Physics*, 21, 447, July 1 1949.
- Goldstein, H. (1980). *Classical mechanics* (2nd ed.). Addison Wesley.
- Griffiths, D. (1987). *Introduction to elementary particles*. New York: Wiley.
- Hawking, S. & Ellis, G. F. R. (1973). *The large scale structure of space-time*. Cambridge University Press.
- Hawking, S. W. (1974). Black hole explosions? *Nature*, 248(5443), 30–31.
- Hetherington, N. S. (1980). Sirius B and the gravitational redshift—an historical review. *Quarterly Journal Royal Astronomical Society*, 21, 246–252.
- Holz, D. E., Hughes, S. A. & Schutz, B. F. (2018). *Measuring cosmic distances with standard sirens*. *Physics Today* 35, December 2018.
- Hoyle, F. (1948). A new model for the expanding universe. *MNRAS*, 108, 372.
- Hubble, E. (1929). A relation between distance and radial velocity among extra-galactic nebulae. *Proceedings of National Academic Sciences*, 15(3), 168–173.
- Hulse, R. A., & Taylor, J. H. (1975). Discovery of a pulsar in a binary system. *Astrophysics Journal*, 195, L51–L53.
- Jackson, J. D. (1999). *Classical electrodynamics*, (3rd ed.). Wiley 1999.
- Kasen, D. et al (2017). *Origin of the heavy elements in binary neutron star mergers from a gravitational wave event*. arxiv.org/abs/1710.05463.
- Kenyon, I. R. (1990). *General relativity*. Oxford University Press.
- Kerr, R. P. (1963). Gravitational field of a spinning mass as an example of algebraically special metrics. *Physical Review Letters*, 11(5), 237–238.
- Kinney, W. H. (2002). *Cosmology, inflation, and the physics of nothing*. [arxiv:astro-ph/0301448](https://arxiv.org/abs/astro-ph/0301448).
- Kirshner, R. P. (2004). Hubble's diagram and cosmic expansion. *Proceedings of National Academic Sciences*, 101(1), 8–13.
- Kofman, L. (1996). *The origin of matter in the universe: Reheating after inflation*. [arxiv:astro-ph/9605155](https://arxiv.org/abs/astro-ph/9605155).
- Kruskal, M. (1960). Maximal extension of Schwarzschild metric. *Physical Review*, 119, 1743.
- Knox, L. and Millea, M. (2019). *The hubble hunter's guide*. [arXiv:1908.03663v2](https://arxiv.org/abs/1908.03663v2).
- Lawrie, I. D. (1990). *A unified grand tour of theoretical physics*. Adam Hilger.
- Le Verrier, U. (1859). Lettre de M. Le Verrier à M. Faye sur la théorie de Mercure et sur le mouvement du périhélie de cette planète. *Comptes rendus hebdomadaires des séances de l'Académie des sciences (Paris)*, 49(1859), 379–383.
- Liddle, A. (2003). *An introduction to modern cosmology* (2nd edn.). Wiley.
- Liddle, A. R. (1999). *An introduction to cosmological inflation*. [arxiv:astro-ph/9901124](https://arxiv.org/abs/astro-ph/9901124).
- LIGO collaboration website, ligo.caltech.edu.
- LIGO Collaboration. (2017). A gravitational-wave standard candle measurement of the Hubble constant. *Nature*, 551, 85.
- Linde, A. D. (2007). *Inflationary cosmology*. arxiv.org/abs/1705.0164.
- Lommen, A. N. (2017). Pulsar timing for gravitational wave detection. *Nature Astronomy*, 1, 809–811.
- Mannheim, P. D. (2011). *Making the case for conformal gravity*. [arXiv:1101.2186](https://arxiv.org/abs/1101.2186), 2011.
- Martins, C. J. A. P., Marinelli, M., Calabrese, M.P., Ramos L. P. (2016). *Real-time cosmography with redshift derivatives*. arxiv.org/abs/1606.07261.
- Milgrom, M. (2014). *MOND theory*. [arXiv:1404.7661](https://arxiv.org/abs/1404.7661).
- Misner, C., Thorne, K., & Wheeler, J. (1973). *Gravitation*. U.S.: W. H. Freeman.

- Moskowitz, C. (2019). *What happened to all the universe's antimatter?* May 23, 2019, scientificamerican.com.
- Narayan, R. (1997). *Lectures on gravitational lensing*. arxiv.org/abs/1907.05922.
- NASA. (2019). Website on the Λ CDM model, containing many original references. lambda.gsfc.nasa.gov/Norton.
- Newcomb S. (1895). *The elements of the four inner planets and the fundamental constants of astronomy. Supplementary American Ephemeris and Nautical Almanac for 1897*, Washington, D.C., Gov. Printing Office, pp. 1–202.
- Ohanian, H.C. & Ruffini, R. (1994). *Gravitation and Spacetime*. Norton.
- Oppenheimer, J. R., & Snyder, H. (1939). On continued gravitational contraction. *Physical Review*, 56, 455.
- Oppenheimer, J. R., & Volkoff, G. M. (1939). On massive neutron cores. *Physical Review*, 55(4), 374–381.
- Pauli, W. (1958). *Theory of relativity*. Pergamon Press, London. This is a translation from an early 1921 encyclopedia article by Pauli, a famous and clear expositions of the theory.
- Peebles, P. J. E. (1965). The black-body radiation content of the universe and the formation of galaxies. *Astrophysics Journal*, 142, 1317.
- Peebles, P. J. E. (1968). Recombination of the primeval plasma. *Astrophysics Journal*, 153, 1.
- Peebles, P. J. E. (1993). *Principles of physical cosmology*. Princeton Press.
- Perlis, S. (1952). *Theory of matrices*. Addison-Wesley. This classic is a clear and self-contained exposition of matrix theory.
- Peskin, M. (2019). *Concepts of elementary particle physics*. Oxford Master Series.
- Petrov, A. Z. (1969). *Einstein spaces*. Pergamon Press.
- Planck, M. (1899). Naturliche Masseinheiten. *Der Koniglich Preussishen Akademie Der Wissenschaften*, 479.
- Planck Collaboration. (2018). Planck 2018 results. VI. Cosmological parameters. arXiv:1807.06209.
- Pound, R. V. (2000). Weighing Photons. *Classical and Quantum Gravity*, 17(12), 2303–2311.
- Quigg, C. (2006). Particles and the standard model. In G. Fraser (Ed.), *The new physics for the twentieth-first century*. Cambridge University Press.
- Randall, L. (2018). What is dark matter? *Nature*, 557, S6–S7.
- Reiss, A. G., Casertaeno, S., Yuan, W., Macri, L. M. & Scolnic, D. (2019). *Large magellanic cloud cepheid standards provide a 1% foundation for the determination of the hubble constant and stronger evidence for physics beyond Λ CDM*. arXiv:1903.07603v2 [astro-ph.CO] Mar 2019.
- Rindler, W. (1969). *Essential relativity*. Van Nostrand and Reinhold.
- Rovelli, C. (2008). Quantum gravity. *Scholarpedia*, 3(5), 7117.
- Rubin, V. (1995). A century of galaxy spectroscopy. *The Astrophysical Journal* 451: 419ff.
- Rubin, V. (1997). *Bright galaxies, dark matters*. Masters of Modern Physics. Woodbury, New York City: Springer Verlag/AIP Press.
- Ruffini, R. & Wheeler, J. A. (1971). *Proceedings of the Conference on Space Physics*. ESRO Paris.
- Sandage, A. R. (1961). The ability of the 200 inch telescope to discriminate between selected world models. *ApJ*, 133(2), 355–392.
- Sard, R. D. (1970). *Relativistic mechanics*. W. A. Benjamin Co., New York, 1970. This presents a simple and careful discussion of the Lorentz transformation in chapters 1 and 2.
- Scardigli, F. (1999). Generalized uncertainty principle in quantum gravity from microscopic black hole gedanken experiment. *Physics Letter B*, 452, 39.
- Schiffer, M. M., Adler, R. J., Mark, J. & Scheffield, C. (1973). Kerr geometry as complexified Schwarzschild geometry. *J. Math. Phys. (N.Y.)*, 14(1), 52–56.
- Schneider, P., Ehlers, J. & Falco, E.E. (1992). *Gravitational lenses*. Springer-Verlag.
- Schutz, B. F. (1986). *Nature*, 323(310).
- Schutz, B. F. (2009). *A first course in general relativity* (2nd ed.). Cambridge University Press.
- Schwartz, H. (1968). *Introduction to special relativity*. McGraw Hill.

- Schwarzschild, K. (1916). *On the gravitational field of a mass point in the Einstein theory (English translation)*. Wiss: Sitzber. Preuss. Akad.
- Shapiro, I. I., et al. (1971). Fourth Test of General Relativity: New Radar Result. *Physical Review Letters*, 26(18), 1132–1135.
- Smoot, G. F. (2006). *Cosmic microwave background radiation anisotropies: their discovery and utilization, nobel lecture 2006*, Nobel Foundation.
- Susskind, L. (2005). *The cosmic landscape: String theory and the illusion of intelligent design*. Little Brown and Company.
- Taylor, E. & Wheeler, J. (1963). *Spacetime physics*. W. H. Freeman.
- Tegmark, M. (2019). *Max Tegmark's CMB data analysis center*. space.mit.edu.
- Thirring, H. Phys. Z., 19, 33; 22 (1921) 29; J. Lense and H. Thirring, Phys. Z., 19, (1918) 156. The English translation can be found in B. Mashhoon, F.W. Hehl and D.S. Theiss, Gen. Rel. Grav., 16 (1984) 711.
- Tolman, R. C. (1939). Static solutions of Einstein's field equations for spheres of fluid. *Physical Review*, 55(4), 364–373.
- Trautman, A. (2006). *Einstein-cartan theory*. arXiv:gr-qc/0606062.
- Vaas, R. (2010). *Multiverse scenarios in cosmology: Classification, cause, challenge, controversy and criticism*. arXiv.org/abs/1001.0726.
- Vessot, R. F. C. et al. (1980). Test of relativistic gravitation with a space-borne hydrogen maser. *Physical Review Letters*, 45(26), 2081–2084.
- von Klüber, H. (1960). Determination of Einstein's light-deflection in the gravitational field of the sun. *Vistas in Astronomy*, 3, Pergamon.
- WMAP. (2010). *Universe 101*. wmap.gsfc.nasa.gov.
- Wagoner, R. V., Fowler, W. A., & Hoyle, F. (1967). On the synthesis of elements at very high temperatures. *Astrophys. J.*, 148.
- Weaver, J.H. (1987). *The world of physics* (Vol. II). Simon and Shuster.
- Weinberg, S. (1972). *Gravitation and cosmology*. Wiley.
- Weinberg, S. (1988). *The first three minutes, updated edition*. New York: Basic Books.
- Weisberg, J. M. & Taylor, J. H. (2005). *The relativistic binary pulsar B1913+16: Thirty years of observations and analysis*. arXiv:astro-ph/0407149.
- Wiki, D. M. *Dark matter*. en.wikipedia.org.
- Wiki GC. *Gravitational collapse*. en.wikipedia.org.
- Wiki NS. *Neutron stars*. en.wikipedia.org.
- Wiki TGR. *Tests of general relativity*. en.wikipedia.org.
- Wiki STEP. *Space tests of the equivalence principle*. en.wikipedia.org.
- Will, C. (1993). *Theory and experiment in gravitational physics revised*. Cambridge: Cambridge University Press.
- Will, C. (2014). *The confrontation between general relativity and experiment* (Springer, 2014). This is an Open Access review article on the Springer link website. It contains an extensive bibliography.
- Zee, A. (1989). *An old man's toy*. New York: MacMillan.
- Zwicky, F. (1933). Die Rotverschiebung von extragalaktischen Nebeln. *Helvetica Physica Acta*, 6..

Index

A

Absolute time, 3–5, 8, 9
Abstract view, 38, 40, 48, 59, 61, 73, 85
Accelerated motion, 11, 23, 25
Accretion disk, 152
Acoustic peaks in CMB, 272
Affine connection, 59–61, 63–66, 70, 73–75, 78, 79, 102, 231
Age of the universe, 236, 239, 244, 248, 252, 262, 270
Arc length, 15, 27, 28, 68–71, 74, 85, 89, 144

B

Basis, 4, 40–44, 46, 48–52, 55, 58, 59, 61, 73, 74, 85, 86, 90, 92, 96, 199, 211, 224, 260
Big bang, 205, 244, 249, 250, 264–266, 269, 271, 277, 281
Binary black holes, 249
Birkhoff theorem, 129
Black hole, 91, 128, 141–145, 147–158, 178, 179, 181, 182, 189, 206, 218, 227, 249, 276, 291, 292, 295, 299–301
Black hole entropy, 156, 158

C

Cartesian, 3, 5, 13, 33, 35, 37, 54, 55, 57, 59–61, 77, 78, 83, 87, 91, 109, 122, 210, 260, 261, 271, 298
Chirp, 176–179, 181, 189, 249
Christoffel connections, 61, 66, 75, 77, 78, 85
Collapse, gravitational, 141

Component, 14, 16, 21, 24, 38, 40–44, 46–51, 59–61, 73, 74, 81, 85–87, 89–91, 102, 104, 110, 113, 115, 116, 119, 122, 123, 127, 138, 140, 151, 157, 163–167, 169, 172, 174, 185–187, 194, 206, 208, 210–212, 224, 231
Condensation, 277, 278
Conservation of energy-momentum, 172, 195, 198, 201
Contravariant, 15, 17, 18, 39–41, 44, 46–48, 60, 78, 82, 83, 89, 113
Coordinates, 4–6, 13–19, 33–52, 54–64, 67, 71–79, 83, 85, 87–92, 100, 102, 103, 105, 109, 110, 112–115, 121–123, 125, 126, 129, 130, 137, 139, 140, 142, 144–147, 151, 157, 160–162, 164, 169, 188, 207–213, 215–217, 220, 223, 228, 235, 238, 257–261, 269–271, 281–283, 298
Cosmic Microwave Background (CMB), 205, 218, 219, 244, 248–250, 263–266, 269–273, 275, 277, 279, 281, 286, 289
Cosmological constant, 193, 198–201, 205, 224, 225, 227, 228, 231, 233, 234, 236, 238, 241, 242, 251, 253–255, 259, 282, 286, 290, 291, 296
Cosmology, 38, 55, 117, 120, 123, 150, 193, 198–200, 203, 205, 211, 216, 223, 224, 227, 238, 241, 244, 247, 250, 252, 257, 270, 272, 277, 287, 290
Covariant, 16–18, 22, 39–41, 44–48, 60, 66, 81–89, 91, 92, 110–114, 117, 122, 160, 198, 285, 297, 298
Critical density, 225, 227, 230, 254, 255, 278

Curvature, 38, 64, 66, 79, 100, 109, 121, 225, 230, 231, 235–237, 239, 240, 251, 254–256, 259, 262, 273, 293, 296
 Curvature parameter, 208, 210, 211, 217, 220, 224–226, 237, 244, 245, 254, 257, 273
 Curvature tensor, 100, 111, 157

D

D'Alembertian operator, 162, 182
 Dark energy, 193, 198–201, 205, 206, 219, 225–227, 231, 234, 241–245, 248, 251, 253, 254, 259, 275, 290
 Dark matter, 157, 201, 205, 206, 226, 227, 232, 238, 248, 250, 251, 254, 275, 278, 291, 292, 295, 299, 301
 Deceleration parameter q , 216, 233, 247, 251
 Decoupling, 251, 252, 266–271, 273, 278, 279, 281, 282
 Deflection of light, 121, 134
 De Sitter universe, 220, 259, 262
 Determinant, metric, 36, 53–55, 57, 88, 91, 126, 140, 284, 297, 298
 Divergence, 87, 88, 91, 92, 116–118, 120, 162, 163, 172, 173, 185, 187, 193–198, 201
 Doppler effect, 10
 Dust, 118, 149, 150, 152, 153, 163, 193–196, 201, 227
 Dust star, 143, 149, 150, 157
 Dust tensor, 119, 120, 193, 196

E

Eddington parameters, 139, 140
 Electromagnetism, 3, 4, 90, 106, 160, 161, 163, 182, 186, 189
 Energy, 19, 21–23, 28, 29, 77–79, 96, 106, 116–118, 120, 141, 148, 151–156, 158, 164, 172, 176, 177, 179, 181, 189, 190, 193–201, 205, 206, 219, 223, 225–229, 241, 249–251, 253, 256, 257, 263–268, 272, 273, 275–279, 282–287, 290, 292–296, 299–302
 Energy-momentum tensor, 117–120, 161, 163, 164, 172–174, 193–200, 223, 224, 285, 298, 299
 Equivalence principle, 64, 98–102, 105, 121, 122, 134, 137, 138
 Euclidian, 221
 Euler-Lagrange equations, 70–72, 76, 77, 79, 129, 130, 168, 297

Event horizon, 153
 Event Horizon Telescope (EHT), 153, 154
 Expansion of the universe, 205, 227, 238, 243, 244, 249, 259, 264, 265, 269, 279
 Exterior derivative, 90
 Extremum curves, 68–71

F

Fictitious forces, 77, 78, 99, 100, 102, 104
 Field theory, 77, 154, 155, 170, 200, 284, 290, 291, 296, 297, 302
 Flat space, 36, 58, 69, 74, 83, 111, 116, 126, 127, 138, 139, 161, 194, 196, 197, 259, 302
 FLRW metric, 207, 211, 212, 214, 216, 221, 223, 224, 231, 257–259, 261, 285
 Four-velocity, 20–24, 119, 120, 194, 211, 224
 Friedmann equation, 224, 234, 250, 296

G

Galaxy, 140, 153, 195, 201, 203–206, 211–220, 226, 227, 232, 238, 242, 244, 245, 247, 248, 250, 258, 261, 263, 275–277, 286, 301
 Galilean transformation, 3, 5, 18
 Gauge transformation, 160, 161, 163, 166, 185, 186
 Generalized Uncertainty Principle (GUP), 293–295, 300, 301
 Geodesic, 59, 64, 66–69, 71, 72, 74, 75, 77–79, 103, 106, 113, 114, 134, 135, 168, 169, 189, 211, 269
 Geodesic system, 64, 75, 113, 114, 160
 Geometric mass, 128, 131, 141, 151, 157
 Gluons, 195, 265, 266, 276, 278, 279
 Godel model universe, 244
 Gravitational redshift, 101, 105, 106, 137, 141
 Gravitational waves, 91, 138, 153, 159, 161, 166, 168–172, 175–177, 179, 180, 182, 184, 186–190, 206, 218, 219, 249–251, 276
 Gravity, 55, 64, 78, 95–98, 100, 102, 106, 109, 110, 115–117, 119–121, 123, 125, 127, 137–139, 141, 148–150, 156, 159, 163, 166, 168, 186, 189, 195, 197, 198, 201, 206, 243, 245, 277, 279, 280, 292–294, 301

H

Halo, dark matter, 226
 Hawking radiation, 153, 156–158, 291, 292, 301
 Homogeneous, 63, 133, 206–210, 220, 244, 279, 285
 Horizon, 145, 151, 238, 257, 259, 262, 270, 279, 281–284, 286–288, 296, 301
 Horizon puzzle, 269, 270, 273, 279, 281–284
 Hubble constant H_0 , 182, 204–206, 214–220, 225, 226, 235, 247–251, 262, 272, 290
 Hubble law, 205, 214, 216
 Hubble length, 287–289
 Hypersphere, 207, 217, 257
 Hypersphere, pseudo-hypersphere, 217, 220

I

Index, 16, 17, 34, 41, 44, 47–50, 56, 72, 73, 84, 91, 113, 114, 116, 122, 123, 167, 174, 194, 223, 231, 298, 299
 Index juggling, 34, 44, 65, 82
 Inflation, 205, 231, 244, 249, 250, 270, 279, 281–290, 293, 295, 296
 Inflaton field, 276, 284–287, 290, 302
 Inner product, 17, 18, 24, 40–42, 44, 64–66, 74, 83, 146
 Invariant, 13–15, 17–24, 36, 38, 40, 41, 45, 46, 53, 55, 69, 77, 85, 88, 89, 91, 92, 121, 134, 160, 238
 Isotropic, 129, 137, 139, 164, 205–210, 220, 244, 269, 270, 279, 281, 285
 Isotropic Schwarzschild metric, 140

K

Kepler, Johannes, 291
 Kerr metric, 148, 150–152

L

Ladder method, 219, 249
 Lagrangian, 70, 76, 77, 79, 129, 130, 134, 135, 168, 284, 285, 297, 298, 302
 Laplacian, 87–89, 91, 92, 163, 285, 298
 Laser Interferometric Gravitational Observatory (LIGO), 138, 180–182, 219, 251
 Last scattering surface, 283
 Length contraction, 8–10
 Lepton, 275, 276, 279, 286
 Levi-Cevita symbol, epsilon, 91, 92

Light cone, 13, 14, 18–20, 27, 145–147, 157, 257–259, 261, 269
 Light, speed, 4–6, 13, 97, 182, 257, 259, 269, 273
 Linearized theory, 129, 139, 159
 Line element, 20, 27, 34–38, 41, 45, 49, 52, 55, 57, 67–71, 77, 79, 91, 103, 126–129, 134, 135, 139, 142, 145, 146, 164, 165, 168, 170, 188, 194, 207, 213, 259, 261
 Lorentz group, 13, 15, 17, 18, 45, 52
 Lorentz transformation, 5, 7–9, 11–13, 15, 17, 18, 22, 23, 39, 97

M

Metric, 14, 16, 18, 28, 29, 34–36, 38, 40, 41, 43, 45, 47–58, 65, 66, 68, 71, 73, 75, 78, 79, 84–88, 91, 92, 102–106, 109–113, 115, 116, 119, 120, 122–130, 138–141, 145, 146, 149–151, 159–161, 163–169, 172, 174, 175, 188, 189, 196, 198, 207–212, 216, 217, 220, 221, 223, 260–262, 285, 297, 298
 Momentum, 19, 21–23, 28, 29, 120, 150, 151, 155, 172, 194–198, 201, 293, 294, 300
 Muon, 9

N

Neutrino, 201, 219, 227, 272, 275, 277–279
 Neutron, 148, 149, 276, 278
 Neutron star, 138, 149, 152, 153, 158, 178, 179, 182, 189, 206, 218, 249, 250, 276
 Newton, 3–5, 78, 95, 96, 98, 102, 117, 170, 196, 197, 201
 Newtonian gravity, 95, 96, 102, 120, 121, 137, 242, 245
 Newtonian mechanics, 67
 N-trads, 49–51
 Nuclei, 153, 275–279
 Nucleon, 250, 275–279
 Null surface, 145, 147, 148, 151

O

Observable universe, 219, 257, 258, 262, 281
 Observational tests, 129, 137, 139, 180, 279, 289
 One-way membrane, 145, 147, 151
 Orbit of a planet, 129, 132, 134

P

Parallel displacement, 64–66, 68, 69, 73, 81, 102, 112, 113
 P-forms, 55, 90, 91
 Planck distribution, 264, 265
 Planck scale, 156, 158, 280, 292, 293, 295, 301
 Plane waves, 164–166, 168, 171, 174, 184, 187
 Poisson's equation, 120, 163, 164, 243
 Polarization, CMB, 165, 169, 170, 175, 180, 185, 250
 Power law inflation, 282, 295, 296
 PPN parameters, 137, 139
 Precession of orbit, 133, 140, 189
 Proper length, 9, 10
 Proper time, 8, 19, 20, 23, 25–28, 68, 103, 142, 144, 148, 169, 194, 210
 Proton, 140, 148, 180, 266, 276–278, 302

Q

Quadrupole formula, 174, 176, 179
 Quarks, 195, 265, 275, 277–279, 286
 Quotient theorem, 46, 47, 85, 111

R

Radiation era, 249, 265–267, 270–272, 275, 277, 281–283, 288
 Rapidity, 11, 12, 25, 26
 Redshift z , 250
 Ricci, 84, 85, 87, 113, 161, 231
 Ricci tensor, 116–119, 123, 126, 140, 157, 161, 231
 Riemann, 33, 34, 37, 38, 52, 59, 61, 64, 66, 73, 74, 100, 111, 113, 116, 118, 122, 123, 157, 161, 220, 231
 Riemann tensor, 109, 111–116, 121–123, 157, 160, 161, 165, 166

S

Scalar, 18, 38–40, 42, 45, 46, 49, 52–54, 67, 81, 83, 88, 90–92, 116, 118, 123, 157, 161, 165, 185, 194, 196, 220, 231, 284, 297
 Scalar field, 83, 279, 284, 285, 296–299, 302
 Schwarzschild metric, 128, 135, 140–142, 151, 152, 157, 158, 242
 Schwarzschild radius, 128, 139–142, 145, 148–150, 153, 155, 157, 158, 242, 300
 Self-parallel curves, 67

Signature, 36, 38, 53, 56, 57, 68, 75, 98, 100, 109, 121, 122, 145, 146, 166
 Slow roll, 285, 286
 Spacetime, 13, 15, 19, 26, 27, 33, 36, 38, 54, 99, 100, 102, 104–106, 109, 118, 121–123, 129, 145, 150, 165, 183, 200, 257, 276, 280, 293–295, 297
 Spectrum of CMB, 205, 250, 264, 279, 286
 Steady state universe, 244
 Stellar evolution, 148

T

Tangent space, 122
 Tangent vector, 67, 68, 74, 86, 89, 146, 147
 Tensor, 14, 15, 18, 28, 29, 33, 34, 36, 44–49, 52, 53, 55, 56, 59, 61–63, 65, 66, 81–88, 90, 92, 109–118, 123, 125, 151, 161, 164, 185, 193, 195–198, 223, 224, 231, 298
 Tetrad, 49–52
 Tidal forces, 100, 121, 170
 Time dilation, 7, 8, 11, 20
 Trajectory, 19, 20, 25–28, 38, 68, 77, 98, 104, 147, 157, 258

U

Uncertainty Principle (UP), 154, 155, 293, 300, 302
 Universe, 9, 11, 117, 118, 145, 150, 152, 156, 182, 193, 195, 198–201, 203, 205–207, 213, 214, 216, 217, 220, 225–230, 233–245, 247–254, 256–267, 269, 270, 272, 273, 275–277, 279–282, 284–293, 295, 296, 301, 302

V

Vacuum, 5, 13, 115–117, 123, 128, 154, 164, 171, 198–200, 225, 230, 254–256, 263, 275, 282, 285, 286, 290, 296
 Vacuum energy density, 250, 255–257, 284, 299
 Vacuum field equations, 117, 125, 129
 Vector, 4, 13, 15–24, 33, 38–52, 58–69, 73, 74, 78, 79, 81–92, 95, 102, 106, 109–113, 146, 161, 164, 165, 184–188, 194
 Vector transplantation, 61, 62, 64–66, 73, 74, 102
 Volume element, 52–55, 88, 91, 92, 145

W

Wave equation, [91](#), [163](#), [165](#), [166](#), [184–186](#)

Weyl theorem, [64](#), [78](#)

White dwarf star, [137](#), [148](#), [149](#)