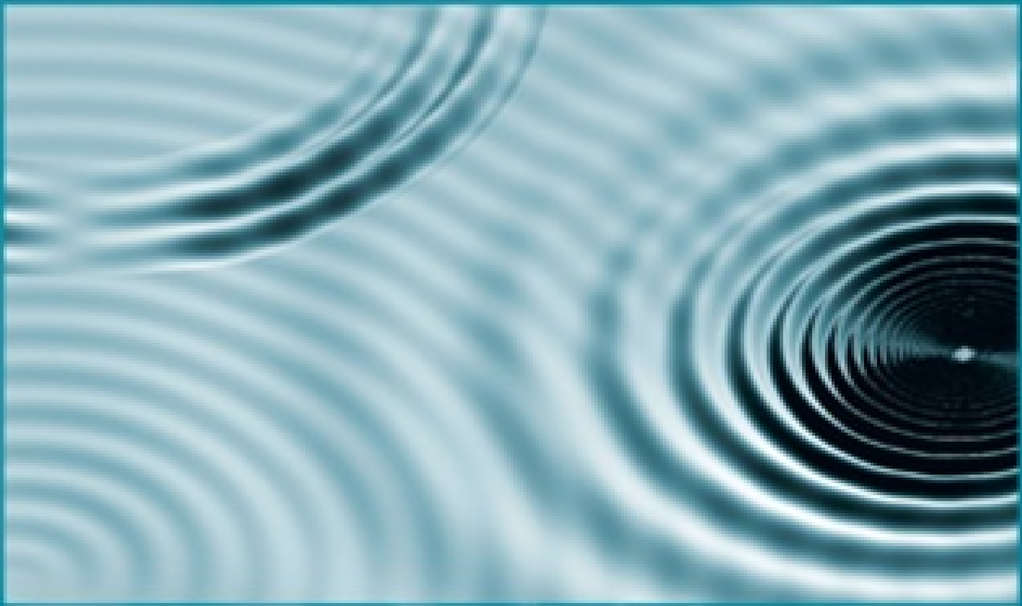
Second Edition

A First Course in GENERAL RELATIVITY



Bernard Schutz

CAMBRIDGE

more information – www.cambridge.org/9780521887052

A First Course in General Relativity

Second Edition

Clarity, readability, and rigor combine in the second edition of this widely used textbook to provide the first step into general relativity for undergraduate students with a minimal background in mathematics.

Topics within relativity that fascinate astrophysical researchers and students alike are covered with Schutz's characteristic ease and authority – from black holes to gravitational lenses, from pulsars to the study of the Universe as a whole. This edition now contains recent discoveries by astronomers that require general relativity for their explanation; a revised chapter on relativistic stars, including new information on pulsars; an entirely rewritten chapter on cosmology; and an extended, comprehensive treatment of modern gravitational wave detectors and expected sources.

Over 300 exercises, many new to this edition, give students the confidence to work with general relativity and the necessary mathematics, whilst the informal writing style makes the subject matter easily accessible. Password protected solutions for instructors are available at www.cambridge.org/Schutz.

Bernard Schutz is Director of the Max Planck Institute for Gravitational Physics, a Professor at Cardiff University, UK, and an Honorary Professor at the University of Potsdam and the University of Hannover, Germany. He is also a Principal Investigator of the GEO600 detector project and a member of the Executive Committee of the LIGO Scientific Collaboration. Professor Schutz has been awarded the Amaldi Gold Medal of the Italian Society for Gravitation.

A First Course in General Relativity

Second Edition

Bernard F. Schutz

Max Planck Institute for Gravitational Physics (Albert Einstein Institute)
and
Cardiff University

CAMBRIDGE UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo, Delhi, Mexico City

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York www.cambridge.org

Information on this title: www.cambridge.org/9780521887052

First published 2009

4th printing 2013

Printed and bound in the United Kingdom by the MPG Books Group *A catalog record for this publication is available from the British Library* ISBN 978-0-52188705-2 hardback

To Siân

Contents

Preface to the second edition

In the 23 years between the first edition of this textbook and the present revision, the field of general relativity has blossomed and matured. Upon its solid mathematical foundations have grown a host of applications, some of which were not even imagined in 1985 when the first edition appeared. The study of general relativity has therefore moved from the periphery to the core of the education of a professional theoretical physicist, and more and more undergraduates expect to learn at least the basics of general relativity before they graduate.

My readers have been patient. Students have continued to use the first edition of this book to learn about the mathematical foundations of general relativity, even though it has become seriously out of date on applications such as the astrophysics of black holes, the detection of gravitational waves, and the exploration of the universe. This extensively revised second edition will, I hope, finally bring the book back into balance and give readers a consistent and unified introduction to modern research in classical gravitation.

The first eight chapters have seen little change. Recent references for further reading have been included, and a few sections have been expanded, but in general the geometrical approach to the mathematical foundations of the theory seems to have stood the test of time. By contrast, the final four chapters, which deal with general relativity in the astrophysical arena, have been updated, expanded, and in some cases completely rewritten.

In Ch. 9, on gravitational radiation, there is now an extensive discussion of detection with interferometers such as LIGO and the planned space-based detector LISA. I have also included a discussion of likely gravitational wave sources, and what we can expect to learn from detections. This is a field that is rapidly changing, and the first-ever direct detection could come at any time. Chapter 9 is intended to provide a durable framework for understanding the implications of these detections.

In Ch. 10, the discussion of the structure of spherical stars remains robust, but I have inserted material on real neutron stars, which we see as pulsars and which are potential sources of detectable gravitational waves.

Chapter 11, on black holes, has also gained extensive material about the astrophysical evidence for black holes, both for stellar-mass black holes and for the supermassive black holes that astronomers have astonishingly discovered in the centers of most galaxies. The discussion of the Hawking radiation has also been slightly amended.

Finally, Ch. 12 on cosmology is completely rewritten. In the first edition I essentially ignored the cosmological constant. In this I followed the prejudice of the time, which assumed that the expansion of the universe was slowing down, even though it had not yet been accurately enough measured. We now believe, from a variety of mutually consistent observations, that the expansion is accelerating. This is probably the biggest challenge to theoretical physics today, having an impact as great on fundamental theories of particle physics as on cosmological questions. I have organized Ch. 12 around this perspective, developing mathematical models of an expanding universe that include the cosmological constant, then discussing in detail how astronomers measure the kinematics of the universe, and finally exploring the way that the physical constituents of the universe evolved after the Big Bang. The roles of inflation, of dark matter, and of dark energy all affect the structure of the universe today, and even our very existence. In this chapter it is possible only to give a brief taste of what astronomers have learned about these issues, but I hope it is enough to encourage readers to go on to learn more.

I have included more exercises in various chapters, where it was appropriate, but I have removed the exercise solutions from the book. They are available now on the website for the book.

The subject of this book remains classical general relativity; apart from a brief discussion of the Hawking radiation, there is no reference to quantization effects. While quantum gravity is one of the most active areas of research in theoretical physics today, there is still no clear direction to point a student who wants to learn how to quantize gravity. Perhaps by the third edition it will be possible to include a chapter on how gravity is quantized!

I want to thank many people who have helped me with this second edition. Several have generously supplied me with lists of misprints and errors in the first edition; I especially want to mention Frode Appel, Robert D’Alessandro, J. A. D. Ewart, Steve Fulling, Toshi Futamase, Ted Jacobson, Gerald Quinlan, and B. Sathyaprakash. Any remaining errors are, of course, my own responsibility. I thank also my editors at Cambridge University Press, Rufus Neal, Simon Capelin, and Lindsay Barnes, for their patience and encouragement. And of course I am deeply indebted to my wife Sian for her generous patience during all the hours, days, and weeks I spent working on this revision.

Preface to the first edition

This book has evolved from lecture notes for a full-year undergraduate course in general relativity which I taught from 1975 to 1980, an experience which firmly convinced me that general relativity is not significantly more difficult for undergraduates to learn than the standard undergraduate-level treatments of electromagnetism and quantum mechanics. The explosion of research interest in general relativity in the past 20 years, largely stimulated by astronomy, has not only led to a deeper and more complete understanding of the theory, it has also taught us simpler, more physical ways of understanding it. Relativity is now in the mainstream of physics and astronomy, so that no theoretical physicist can be regarded as broadly educated without some training in the subject. The formidable reputation relativity acquired in its early years (Interviewer: 'Professor Eddington, is it true that only three people in the world understand Einstein's theory?' Eddington: 'Who is the third?') is today perhaps the chief obstacle that prevents it being more widely taught to theoretical physicists. The aim of this textbook is to present general relativity at a level appropriate for undergraduates, so that the student will understand the basic physical concepts and their experimental implications, will be able to solve elementary problems, and will be well prepared for the more advanced texts on the subject.

In pursuing this aim, I have tried to satisfy two competing criteria: first, to assume a minimum of prerequisites; and, second, to avoid watering down the subject matter. Unlike most introductory texts, this one does not assume that the student has already studied electromagnetism in its manifestly relativistic formulation, the theory of electromagnetic waves, or fluid dynamics. The necessary fluid dynamics is developed in the relevant chapters. The main consequence of not assuming a familiarity with electromagnetic waves is that gravitational waves have to be introduced slowly: the wave equation is studied from scratch. A full list of prerequisites appears below.

The second guiding principle, that of not watering down the treatment, is very subjective and rather more difficult to describe. I have tried to introduce differential geometry fully, not being content to rely only on analogies with curved surfaces, but I have left out subjects that are not essential to general relativity at this level, such as nonmetric manifold theory, Lie derivatives, and fiber bundles.[1] I have introduced the full nonlinear field equations, not just those of linearized theory, but I solve them only in the plane and spherical cases, quoting and examining, in addition, the Kerr solution. I study gravitational waves mainly in the linear approximation, but go slightly beyond it to derive the energy in the waves and the reaction effects in the wave emitter. I have tried in each topic to supply enough foundation for the student to be able to go to more advanced treatments without having to start over again at the beginning.

The first part of the book, up to Ch. 8, introduces the theory in a sequence that is typical of many treatments: a review of special relativity, development of tensor analysis and continuum physics in special relativity, study of tensor calculus in curvilinear coordinates in Euclidean and Minkowski spaces, geometry of curved manifolds, physics in a curved spacetime, and finally the field equations. The remaining four chapters study a few topics that I have chosen because of their importance in modern astrophysics. The chapter on gravitational radiation is more detailed than usual at this level because the observation of gravitational waves may be one of the most significant developments in astronomy in the next decade. The chapter on spherical stars includes, besides the usual material, a useful family of exact compressible solutions due to Buchdahl. A long chapter on black holes studies in some detail the physical nature of the horizon, going as far as the Kruskal coordinates, then exploring the rotating (Kerr) black hole, and concluding with a simple discussion of the Hawking effect, the quantum mechanical emission of radiation by black holes. The concluding chapter on cosmology derives the homogeneous and isotropic metrics and briefly studies the physics of cosmological observation and evolution. There is an appendix summarizing the linear algebra needed in the text, and another appendix containing hints and solutions for selected exercises. One subject I have decided not to give as much prominence to, as have other texts traditionally, is experimental tests of general relativity and of alternative theories of gravity. Points of contact with experiment are treated as they arise, but systematic discussions of tests now require whole books (Will 1981).[2] Physicists today have far more confidence in the validity of general relativity than they had a decade or two ago, and I believe that an extensive discussion of alternative theories is therefore almost as out of place in a modern elementary text on gravity as it would be in one on electromagnetism.

The student is assumed already to have studied: special relativity, including the Lorentz transformation and relativistic mechanics; Euclidean vector calculus; ordinary and simple partial differential equations; thermodynamics and hydrostatics; Newtonian gravity (simple stellar structure would be useful but not essential); and enough elementary quantum mechanics to know what a photon is.

The notation and conventions are essentially the same as in Misner *et al.*, *Gravitation* (W. H. Freeman 1973), which may be regarded as one possible follow-on text after this one. The physical point of view and development of the subject are also inevitably influenced by that book, partly because Thorne was my teacher and partly because *Gravitation* has become such an influential text. But because I have tried to make the subject accessible to a much wider audience, the style and pedagogical method of the present book are very different.

Regarding the use of the book, it is designed to be studied sequentially as a whole, in a one-year course, but it can be shortened to accommodate a half-year course. Half-year courses probably should aim at restricted goals. For example, it would be reasonable to aim to teach gravitational waves and black holes in half a year to students who have already studied electromagnetic waves, by carefully skipping some of Chs. 1–3 and most of Chs. 4, 7, and 10. Students with preparation in special relativity and fluid dynamics could learn stellar structure and cosmology in half a year, provided they could go quickly through the first four chapters and then skip Chs. 9 and 11. A graduate-level course can, of course, go much more quickly, and it should be possible to cover the whole text in half a year.

Each chapter is followed by a set of exercises, which range from trivial ones (filling in missing steps in the body of the text, manipulating newly introduced mathematics) to advanced problems that considerably extend the discussion in the text. Some problems require programmable calculators or computers. I cannot overstress the importance of doing a selection of problems. The easy and medium-hard ones in the early chapters give essential practice, without which the later chapters will be much less comprehensible. The medium-hard and hard problems of the later chapters are a test of the student's understanding. It is all too common in relativity for students to find the conceptual framework so interesting that they relegate problem solving to second place. Such a separation is false and dangerous: a student who can't solve problems of reasonable difficulty doesn't really understand the concepts of the theory either. There are generally more problems than one would expect a student to solve; several chapters have more than 30. The teacher will have to select them judiciously. Another rich source of problems is the *Problem Book in Relativity and Gravitation*, Lightman *et al.* (Princeton University Press 1975).

I am indebted to many people for their help, direct and indirect, with this book. I would like especially to thank my undergraduates at University College, Cardiff, whose enthusiasm for the subject and whose patience with the inadequacies of the early lecture notes encouraged me to turn them into a book. And I am certainly grateful to Suzanne Ball, Jane Owen, Margaret Vallender, Pranoat Priesmeyer, and Shirley Kemp for their patient typing and retyping of the successive drafts.

The treatment here is therefore different in spirit from that in my book *Geometrical Methods of Mathematical Physics* (Cambridge University Press 1980*b*), which may be used to supplement this one.

The revised second edition of this classic work is Will (1993).

3

Tensor analysis in special relativity

3.1 The metric tensor

Consider the representation of two vectors $\vec{A}$ and $\vec{B}$ on the basis $\{\vec{e}_\alpha\}$ of some frame $\mathcal{O}$: $\vec{A} = A^\alpha \vec{e}_\alpha, \qquad \vec{B} = B^\beta \vec{e}_\beta.$

Their scalar product is

$$\vec{A} \cdot \vec{B} = (A^\alpha \vec{e}_\alpha) \cdot (B^\beta \vec{e}_\beta).$$

(Note the importance of using *different* indices α and β to distinguish the first summation from the second.) Following Exer. 34, § 2.9, we can rewrite this as $\vec{A} \cdot \vec{B} = A^\alpha B^\beta (\vec{e}_\alpha \cdot \vec{e}_\beta),$

which, by Eq. (2.27), is

$$\vec{A} \cdot \vec{B} = A^\alpha B^\beta \eta_{\alpha\beta}. \tag{3.1}$$

This is a *frame-invariant* way of writing

$$-A^0B^0 + A^1B^1 + A^2B^2 + A^3B^3.$$

The numbers $\eta_{\alpha\beta}$ are called ‘components of the metric tensor’. We will justify this name later. Right now we observe that they essentially give a ‘rule’ for associating with two vectors $\vec{A}$ and $\vec{B}$ a single *number*, which we call their scalar product. The rule is that the number is the double sum $A^\alpha B^\beta \eta_{\alpha\beta}$. Such a rule is at the heart of the meaning of ‘tensor’, as we now discuss.

3.2 Definition of tensors

We make the following definition of a tensor:

A tensor of type $\binom{0}{N}$ is a function of N vectors into the real numbers, which is linear in each of its N

arguments.

Let us see what this definition means. For the moment, we will just accept the notation $\binom{0}{N}$; its justification will come later in this chapter. The rule for the scalar product, Eq. (3.1), satisfies our definition of a $\binom{0}{2}$ tensor. It is a rule which takes two vectors, $\vec{A}$ and $\vec{B}$, and 2 produces a single real number $\vec{A} \cdot \vec{B}$. To say that it is linear in its arguments means what is proved in Exer. 34, § 2.9. Linearity on the first argument means

$$\left.\begin{aligned} (\alpha\vec{A}) \cdot \vec{B} &= \alpha(\vec{A} \cdot \vec{B}), \\ \text{and} \qquad & \\ (\vec{A} + \vec{B}) \cdot \vec{C} &= \vec{A} \cdot \vec{C} + \vec{B} \cdot \vec{C}, \end{aligned}\right\} \tag{3.2}$$

while linearity on the second argument means

$$\begin{aligned} \vec{A} \cdot (\beta\vec{B}) &= \beta(\vec{A} \cdot \vec{B}), \\ \vec{A} \cdot (\vec{B} + \vec{C}) &= \vec{A} \cdot \vec{B} + \vec{A} \cdot \vec{C}. \end{aligned}$$

This definition of linearity is of central importance for tensor algebra, and the student should study it carefully.

To give concreteness to this notion of the dot product being a tensor, we introduce a name and notation for it. We let **g** be the *metric tensor* and write, by definition,

$$\mathbf{g}(\vec{A}, \vec{B}) := \vec{A} \cdot \vec{B}. \tag{3.3}$$

Then we regard **g**(,) as a function which can take two arguments, and which is linear in that

$$\mathbf{g}(\alpha\vec{A} + \beta\vec{B}, \vec{C}) = \alpha\, \mathbf{g}(\vec{A}, \vec{C}) + \beta\, \mathbf{g}(\vec{B}, \vec{C}), \tag{3.4}$$

and similarly for the second argument. The value of **g** on two arguments, denoted by $\mathbf{g}(\vec{A}, \vec{B})$, is their dot product, a real number.

Notice that the definition of a tensor does not mention components of the vectors. A tensor must be a rule which gives the same real number independently of the reference frame in which the vectors' components are calculated. We showed in the previous chapter that Eq. (3.1) satisfies this requirement. This enables us to regard a tensor as a function of the vectors themselves rather than of their components, and this can sometimes be helpful conceptually.

Notice that an ordinary function of position, $f\ (t, x, y, z)$, is a real-valued function of no vectors at all. It is therefore classified as a $\binom{0}{0}$ tensor.

Aside on the usage of the term ‘function’

The most familiar notion of a function is expressed in the equation

$$y = f(x),$$

where *y* and *x* are real numbers. But this can be written more precisely as: *f* is a ‘rule’ (called a mapping) which associates a real number (symbolically called *y*, above) with another real number, which is the argument of *f* (symbolically called *x*, above). The function itself is *not* *f* (*x*), since *f* (*x*) is *y*, which is a real number called the ‘value’ of the function. The function itself is *f*, which we can write as *f* ( ) in order to show that it has one argument. In algebra this seems like hair-splitting since we unconsciously think of *x* and *y* as two things at once: they are, on the one hand, specific real numbers and, on the other hand, *names* for general and arbitrary real numbers. In tensor calculus we will make this distinction explicit: $\vec{A}$ and $\vec{B}$ are *specific* vectors, $\vec{A} \cdot \vec{B}$ is a specific real number, and **g** is the name of the function that associates $\vec{A} \cdot \vec{B}$ with $\vec{A}$ and $\vec{B}$.

Components of a tensor

Just like a vector, a tensor has components. They are defined as:

The components in a frame $\mathcal{O}$ of a tensor of type $\binom{0}{N}$ are the values of the function when its arguments are the basis vectors $\{\vec{e}_\alpha\}$ of the frame $\mathcal{O}$.

Thus we have the notion of components as frame-dependent numbers (frame-dependent because the basis refers to a specific frame). For the metric tensor, this gives the components as

$$\mathbf{g}(\vec{e}_\alpha,\ \vec{e}_\beta) = \vec{e}_\alpha \cdot \vec{e}_\beta = \eta_{\alpha\beta}. \tag{3.5}$$

So the matrix $\eta_{\alpha\beta}$ that we introduced before is to be thought of as an array of the components of **g** on the basis. In another basis, the components could be different. We will have many more examples of this later. First we study a particularly important class of tensors.

3.3 The $\binom{0}{1}$ tensors: one-forms

A tensor of the type $\binom{0}{1}$ is called a covector, a covariant vector, or a one-form. Often these names are used interchangeably, even in a single text-book or reference.

General properties

Let an arbitrary one-form be called $\tilde{p}$. (We adopt the notation that $\tilde{}$ above a symbol denotes a one-form, just as $\vec{}$ above a symbol denotes a vector.) Then $\tilde{p}$, supplied with one vector argument, gives a real number: $\tilde{p}(\vec{A})$ is a real number. Suppose $\tilde{q}$ is another one-form. Then we can define

$$\left.\begin{aligned}\tilde{s} &= \tilde{p} + \tilde{q},\\ \tilde{r} &= \alpha\tilde{p},\end{aligned}\right\} \tag{3.6a}$$

to be the one-forms that take the following values for an argument $\vec{A}$:

$$\left.\begin{aligned}\tilde{s}(\vec{A}) &= \tilde{p}(\vec{A}) + \tilde{q}(\vec{A}),\\ \tilde{r}(\vec{A}) &= \alpha\tilde{p}(\vec{A}).\end{aligned}\right\} \tag{3.6b}$$

With these rules, the set of all one-forms satisfies the axioms for a vector space, which accounts for their other names. This space is called the 'dual vector space' to distinguish it from the space of all vectors such as $\vec{A}$.

When discussing vectors we relied heavily on components and their transformations. Let us look at those of $\tilde{p}$. The components of $\tilde{p}$ are called p_α:

$$p_\alpha := \tilde{p}(\vec{e}_\alpha). \tag{3.7}$$

Any component with a single lower index is, by convention, the component of a one-form; an upper index denotes the component of a vector. In terms of components, $\tilde{p}(\vec{A})$ is

$$\begin{aligned}\tilde{p}(\vec{A}) &= \tilde{p}(A^\alpha\vec{e}_\alpha)\\ &= A^\alpha\tilde{p}(\vec{e}_\alpha),\\ \tilde{p}(\vec{A}) &= A^\alpha p_\alpha.\end{aligned} \tag{3.8}$$

The second step follows from the linearity which is the heart of the definition we gave of a tensor. So the real number $\tilde{p}(\vec{A})$ is easily found to be the sum $A^0p_0 + A^1p_1 + A^2p_2 + A^3p_3$. Notice that *all* terms have plus signs: this operation is called *contraction* of $\vec{A}$ and $\tilde{p}$, and is *more* fundamental in tensor analysis than the scalar product because it can be performed between any one-form and vector without reference to other tensors. We have seen that two vectors cannot make a scalar (their dot product) without the help of a third tensor, the metric.

The components of $\tilde{p}$ on a basis $\{\vec{e}_{\bar{\beta}}\}$ are

$$\begin{aligned} p_{\bar{\beta}} &:= \tilde{p}(\vec{e}_{\bar{\beta}}) = \tilde{p}(\Lambda^{\alpha}{}_{\bar{\beta}}\vec{e}_{\alpha}) \\ &= \Lambda^{\alpha}{}_{\bar{\beta}}\tilde{p}(\vec{e}_{\alpha}) = \Lambda^{\alpha}{}_{\bar{\beta}}p_{\alpha}. \end{aligned} \tag{3.9}$$

Comparing this with

$$\vec{e}_{\bar{\beta}} = \Lambda^{\alpha}{}_{\bar{\beta}}\,\vec{e}_{\alpha},$$

we see that components of one-forms transform in exactly the same manner as basis vectors and in the opposite manner to components of vectors. By 'opposite', we mean using the inverse transformation. This use of the inverse guarantees that $A^{\alpha}p_{\alpha}$ is frame independent for any vector $\vec{A}$ and one-form $\tilde{p}$. This is such an important observation that we shall prove it explicitly:

$$A^{\bar{\alpha}}p_{\bar{\alpha}} = (\Lambda^{\bar{\alpha}}{}_{\beta}A^{\beta})(\Lambda^{\mu}{}_{\bar{\alpha}}p_{\mu}), \tag{3.10a}$$

$$= \Lambda^{\mu}{}_{\bar{\alpha}}\Lambda^{\bar{\alpha}}{}_{\beta}A^{\beta}p_{\mu}, \tag{3.10b}$$

$$= \delta^{\mu}{}_{\beta}A^{\beta}p_{\mu}, \tag{3.10c}$$

$$= A^{\beta}p_{\beta}. \tag{3.10d}$$

(This is the same way in which the vector $A^{\alpha}\vec{e}_{\alpha}$ is kept frame independent.) This inverse transformation gives rise to the word 'dual' in 'dual vector space'. The property of transforming *with* basis vectors gives rise to the *co* in 'covariant vector' and its shorter form 'covector'. Since components of ordinary vectors transform oppositely to basis vectors (in order to keep $A^{\beta}\vec{e}_{\beta}$ frame independent), they are often called 'contravariant' vectors. Most of these names are old-fashioned; 'vectors' and 'dual vectors' or 'one-forms' are the modern names. The reason that 'co' and 'contra' have been abandoned is that they mix up two very different things: the transformation of a basis is the expression of *new* vectors in terms of *old* ones; the transformation of components is the expression of the *same* object in terms of the new basis. It is important for the student to be sure of these distinctions before proceeding further.

Basis one-forms

Since the set of all one-forms is a vector space, we can use any set of four linearly independent one-forms as a basis. (As with any vector space, one-forms are said to be linearly independent if no nontrivial linear combination equals the zero one-form. The zero one-form is the one whose value on any vector is zero.) However, in the previous section we have already used the basis vectors $\{\vec{e}_\alpha\}$ to define the components of a one-form. This suggests that we should be able to use the basis vectors to define an associated one-form basis $\{\tilde{\omega}^\alpha, \alpha = 0, \ldots, 3\}$, which we shall call the basis *dual* to $\{\vec{e}_\alpha\}$, upon which a one-form has the components defined above. That is, we want a set $\{\tilde{\omega}^\alpha\}$ such that

$$\tilde{p} = p_\alpha \tilde{\omega}^\alpha. \tag{3.11}$$

(Notice that using a raised index on $\tilde{\omega}^\alpha$ permits the summation convention to operate.) The $\{\tilde{\omega}^\alpha\}$ are *four distinct* one-forms, just as the $\{\vec{e}_\alpha\}$ are four distinct vectors. This equation must imply Eq. (3.8) for any vector $\vec{A}$ and one-form $\tilde{p}$: $\tilde{p}(\vec{A}) = p_\alpha A^\alpha$.

But from Eq. (3.11) we get

$$\begin{aligned}\tilde{p}(\vec{A}) &= p_\alpha \tilde{\omega}^\alpha(\vec{A}) \\ &= p_\alpha \tilde{\omega}^\alpha(A^\beta \vec{e}_\beta) \\ &= p_\alpha A^\beta \tilde{\omega}^\alpha(\vec{e}_\beta).\end{aligned}$$

(Notice the use of β as an index in the second line, in order to distinguish its summation from the one on α.) Now, this final line can only equal $p_\alpha A^\alpha$ for all A^β and p_α if

$$\tilde{\omega}^\alpha(\vec{e}_\beta) = \delta^\alpha{}_\beta. \tag{3.12}$$

Comparing with Eq. (3.7), we see that this equation gives the βth component of the αth basis one-form. It therefore *defines* the αth basis one-form. We can write out these components as

$$\begin{aligned}\tilde{\omega}^0 &\underset{\mathcal{O}}{\rightarrow} (1, 0, 0, 0), \\ \tilde{\omega}^1 &\underset{\mathcal{O}}{\rightarrow} (0, 1, 0, 0), \\ \tilde{\omega}^2 &\underset{\mathcal{O}}{\rightarrow} (0, 0, 1, 0), \\ \tilde{\omega}^3 &\underset{\mathcal{O}}{\rightarrow} (0, 0, 0, 1).\end{aligned}$$

It is important to understand two points here. One is that Eq. (3.12) defines the basis $\{\tilde{\omega}^\alpha\}$ in terms of the basis $\{\vec{e}_\beta\}$. The vector basis induces a unique and convenient one-form basis. This is not the only possible one-form basis, but it is so useful to have the relationship, Eq. (3.12), between the bases that we will always use it. The relationship, Eq. (3.12), is between the two bases, not between individual pairs, such as $\tilde{\omega}^0$ and $\vec{e}_0$. That is, if we change $\vec{e}_0$, while leaving $\vec{e}_1$, $\vec{e}_2$, and $\vec{e}_3$ unchanged, then in general this induces changes not only in $\tilde{\omega}^0$ but also in $\tilde{\omega}^1$, $\tilde{\omega}^2$, and $\tilde{\omega}^3$. The second point to understand is that, although we can describe both vectors and one-forms by giving a set of four components, their geometrical significance is very different. The student should not lose sight of the fact that the components tell only part of the story. The basis contains the rest of the information. That is, a set of numbers (0, 2, −1, 5) alone does not define anything; to make it into something, we must say whether these are components on a vector basis or a one-form basis and, indeed, which of the infinite number of possible bases is being used.

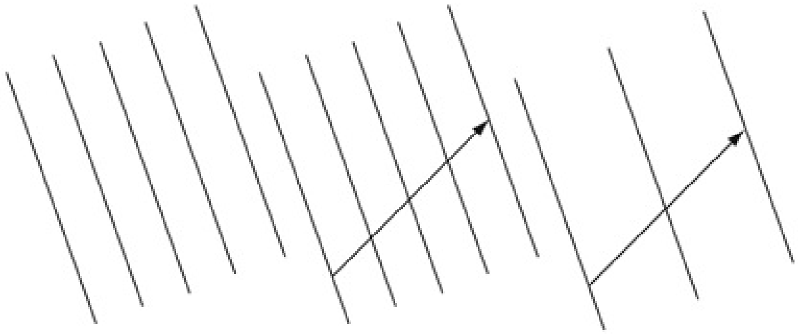
It remains to determine how $\{\tilde{\omega}^\alpha\}$ transforms under a change of basis. That is, each frame has its own unique set $\{\tilde{\omega}^\alpha\}$; how are those of two frames related? The derivation here is analogous to that for the basis vectors. It leads to the only equation we can write down with the indices in their correct positions:

$$\tilde{\omega}^{\bar{\alpha}} = \Lambda^{\bar{\alpha}}{}_{\beta}\tilde{\omega}^{\beta}. \tag{3.13}$$

This is the same as for components of a vector, and opposite that for components of a one-form.

Picture of a one-form

For vectors we usually imagine an arrow if we need a picture. It is helpful to have an image of a one-form as well. First of all, it is not an arrow. Its picture must reflect the fact that it maps vectors into real numbers. A vector itself does not automatically map another vector into a real number. To do this it needs a metric tensor to define the scalar product. With a different metric, the *same* two vectors will produce a *different* scalar product. So two vectors by themselves don’t give a number. We need a picture of a one-form which doesn’t depend on any other tensors having been defined. The one generally used by mathematicians is shown in Fig. 3.1. The one-form consists of a series of surfaces. The ‘magnitude’ of it is given by the spacing between the surfaces: the larger the spacing the *smaller* the magnitude. In this picture, the number produced when a one-form acts on a vector is the number of surfaces that the arrow of the vector pierces. So the closer their spacing, the larger the number (compare (*b*) and (*c*) in Fig. 3.1). In a four-dimensional space, the surfaces are three-dimensional. The one-form doesn't define a unique direction, since it is not a vector. Rather, it defines a way of 'slicing' the space. In order to justify this picture we shall look at a particular one-form, the gradient.



(a) (b) (c)

Figure 3.1 (a) The picture of one-form complementary to that of a vector as an arrow. (b) The value of a one-form on a given vector is the number of surfaces the arrow pierces. (c) The value of a smaller one-form on the same vector is a smaller number of surfaces. The larger the one-form, the more 'intense' the slicing of space in its picture.

Gradient of a function is a one-form

Consider a scalar field $\phi(\vec{x})$ defined at every event $\vec{x}$. The world line of some particle (or person) encounters a value of ϕ at each event on it (see Fig. 3.2), and this value changes from event to event. If we label (parametrize) each point on the curve by the value of proper time τ along it (i.e. the reading of a clock moving on the line), then we can express the coordinates of events on the curve as functions of τ :

$$[t = t(\tau),\ x = x(\tau),\ y = y(\tau),\ z = z(\tau)].$$

The four-velocity has components

$$\vec{U} \rightarrow \left(\frac{\mathrm{d}t}{\mathrm{d}\tau}, \frac{\mathrm{d}x}{\mathrm{d}\tau}, \ldots\right).$$

Since ϕ is a function of *t*, *x*, *y*, and *z*, it is implicitly a function of τ on the curve: $\phi(\tau) = \phi[t(\tau),\ x(\tau),\ y(\tau),\ z(\tau)]$,

and its rate of change on the curve is

$$
\begin{aligned}
\frac{d\phi}{d\tau} &= \frac{\partial\phi}{\partial t}\frac{dt}{d\tau} + \frac{\partial\phi}{\partial x}\frac{dx}{d\tau} + \frac{\partial\phi}{\partial y}\frac{dy}{d\tau} + \frac{\partial\phi}{\partial z}\frac{dz}{d\tau} \\
&= \frac{\partial\phi}{\partial t}U^t + \frac{\partial\phi}{\partial x}U^x + \frac{\partial\phi}{\partial y}U^y + \frac{\partial\phi}{\partial z}U^z.
\end{aligned} \tag{3.14}
$$

It is clear from this that in the last equation we have devised a means of producing from the vector $\vec{U}$ the number $d\phi/d\tau$ that represents the rate of change of ϕ on a curve on which $\vec{U}$ is the tangent. This number $d\phi/d\tau$ is clearly a linear function of $\vec{U}$, so we have defined a one-form.

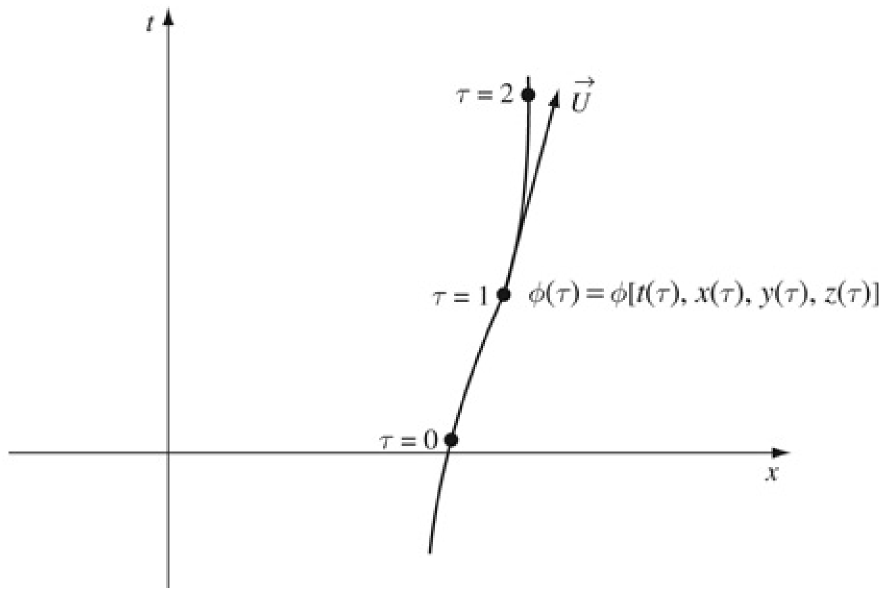


Figure 3.2 A world line parametrized by proper time τ, and the values $\phi(\tau)$ of the scalar field $\phi(t, x, y, z)$ along it.

By comparison with Eq. (3.8), we see that this one-form has components $(\partial\phi/\partial t, \partial\phi/\partial x, \partial\phi/\partial y, \partial\phi/\partial z)$. This one-form is called the *gradient* of ϕ, denoted by $\tilde{d}\phi$:

$$
\tilde{d}\phi \underset{\mathcal{O}}{\rightarrow} \left(\frac{\partial\phi}{\partial t}, \frac{\partial\phi}{\partial x}, \frac{\partial\phi}{\partial y}, \frac{\partial\phi}{\partial z}\right). \tag{3.15}
$$

It is clear that the gradient fits our definition of a one-form. We will see later how it comes about that the gradient is usually introduced in three-dimensional vector calculus as a vector.

The gradient enables us to justify our picture of a one-form. In Fig. 3.3 we have drawn part of a topographical map, showing contours of equal elevation. If h is the elevation, then the gradient $\tilde{d}h$ is clearly largest in an area such as $A,$

where the lines are closest together, and smallest near B, where the lines are spaced far apart. Moreover, suppose we wanted to know how much elevation a walk between two points would involve. We would lay out on the map a line (vector $\Delta\vec{x}$) between the points. Then the number of contours the line crossed would give the change in elevation. For example, line 1 crosses $1\frac{1}{2}$ contours, while 2 crosses two contours. Line 3 starts near 2 but goes in a different direction, winding up only $\frac{1}{2}$ contour higher. But these numbers are just Δh, which is the contraction of $\tilde{d}h$ with $\Delta\vec{x}$: $\Delta h = \sum_i(\partial h/\partial x^i)\Delta x^i$ or the *value* of $\tilde{d}h$ on $\Delta\vec{x}$ (see Eq. (3.8)).

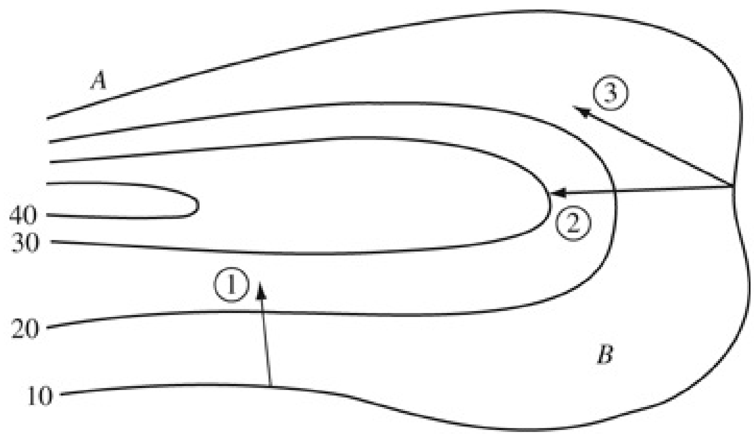


Figure 3.3 A topographical map illustrates the gradient one-form (local contours of constant elevation). The change of height along any trip (arrow) is the number of contours crossed by the arrow.

Therefore, a one-form is represented by a series of surfaces (Fig. 3.4), and its contraction with a vector $\vec{V}$ is the number of surfaces $\vec{V}$ crosses. The closer the surfaces, the larger $\tilde{\omega}$. Properly, just as a vector is straight, the one-form's surfaces are straight and parallel. This is because we deal with one-forms at a point, not over an extended region: 'tangent' one-forms, in the same sense as tangent vectors.

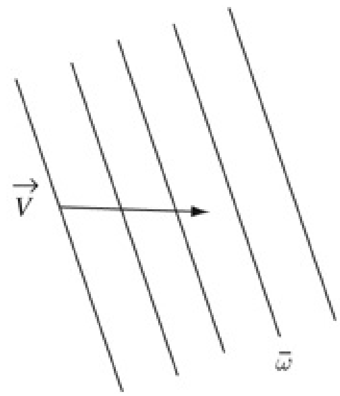


Figure 3.4 The value $\tilde{\omega}(\vec{V})$ is 2.5.

These pictures show why we in general *cannot* call a gradient a vector. We would like to identify the *vector* gradient as that vector pointing 'up' the slope, *i.e.* in such a way that it crosses the greatest number of contours per unit length. The key phrase is 'per unit length'. If there is a metric, a measure of distance in the space, then a vector *can* be associated with a gradient. But the metric must intervene here in order to produce a vector. Geometrically, on its own, the gradient is a one-form.

Let us be sure that Eq. (3.15) is a consistent definition. How do the components transform? For a one-form we must have

$$(\tilde{\mathrm{d}}\phi)_{\bar{\alpha}} = \Lambda^{\beta}{}_{\bar{\alpha}}(\tilde{\mathrm{d}}\phi)_{\beta}. \tag{3.16}$$

But we know how to transform partial derivatives:

$$\frac{\partial \phi}{\partial x^{\bar{\alpha}}} = \frac{\partial \phi}{\partial x^{\beta}} \frac{\partial x^{\beta}}{\partial x^{\bar{\alpha}}},$$

which means

$$(\tilde{\mathrm{d}}\phi)_{\bar{\alpha}} = \frac{\partial x^{\beta}}{\partial x^{\bar{\alpha}}}(\tilde{\mathrm{d}}\phi)_{\beta}. \tag{3.17}$$

Are Eqs. (3.16) and (3.17) consistent? The answer, of course, is yes. The reason: since $x^{\beta} = \Lambda^{\beta}{}_{\bar{\alpha}}\, x^{\bar{\alpha}}$,

and since $\Lambda^{\beta}{}_{\bar{\alpha}}$ are just constants, then

---

$$\partial x^{\beta}/\partial x^{\bar{\alpha}} = \Lambda^{\beta}{}_{\bar{\alpha}}. \tag{3.18}$$

This identity is fundamental. Components of the gradient transform according to the *inverse* of the components of vectors. So the gradient is the 'archetypal' one-form.

Notation for derivatives

From now on we shall employ the usual subscripted notation to indicate derivatives:

$$\frac{\partial \phi}{\partial x} := \phi_{,x}$$

and, more generally,

$$\frac{\partial \phi}{\partial x^\alpha} := \phi_{,\alpha}. \tag{3.19}$$

Note that the index α appears as a superscript in the denominator of the left-hand side of Eq. (3.19) and as a subscript on the right-hand side. As we have seen, this placement of indices is consistent with the transformation properties of the expression.

In particular, we have

$$x^\alpha{}_{,\beta} \equiv \delta^\alpha{}_\beta,$$

which we can compare with Eq. (3.12) to conclude that

$$\tilde{\mathrm{d}}x^\alpha := \tilde{\omega}^\alpha. \tag{3.20}$$

This is a useful result, that the basis one-form is just $\tilde{\mathrm{d}}x^\alpha$. We can use it to write, for any function *f*, $\tilde{\mathrm{d}}f = \frac{\partial f}{\partial x^\alpha}\tilde{\mathrm{d}}x^\alpha.$

This looks very much like the physicist's 'sloppy-calculus' way of writing differentials or infinitesimals. The notation $\tilde{\mathrm{d}}$ has been chosen partly to suggest this comparison, but this choice makes it doubly important for the student to avoid confusion on this point. The object $\tilde{\mathrm{d}}f$ is a tensor, not a small increment in *f* ; it *can* have a small ('infinitesimal') value if it is contracted with a small vector.

Normal one-forms

Like the gradient, the concept of a normal vector – a vector orthogonal to a surface – is one which is more naturally replaced by that of a normal one-form. For a normal vector to be defined we need to have a scalar product: the normal vector must be orthogonal to all vectors tangent to the surface. This can be defined only by using the metric tensor. But a normal one-form can be defined without reference to the metric. A one-form is said to be normal to a surface if its value is zero on every vector tangent to the surface. If the surface is closed and divides spacetime into an 'inside' and 'outside', a normal is said to be an *outward* normal one-form if it is a normal one-form and its value on vectors which point outwards from the surface is positive. In Exer. 13, § 3.10, we prove that $\tilde{d}f$ is normal to surfaces of constant f.

3.4 The $\binom{0}{2}$ tensors

Tensors of type $\binom{0}{2}$ have two vector arguments. We have encountered the metric tensor already, but the simplest of this type is the product of two one-forms, formed according to the following rule: if $\tilde{p}$ and $\tilde{q}$ are one-forms, then $\tilde{p} \otimes \tilde{q}$ is the $\binom{0}{2}$ tensor which, when supplied with vectors $\vec{B}$ as arguments, produces the number $\tilde{p}(\vec{A})\ \tilde{q}(\vec{B})$, *i.e.* just the product of the numbers produced by the $\binom{0}{1}$ tensors. The symbol ⊗ is called an 'outer product sign' and is a formal notation to show how the $\binom{0}{2}$ tensor is formed from the one-forms. Notice that ⊗ is *not* commutative: $\tilde{p} \otimes \tilde{q}$ and $\tilde{q} \otimes \tilde{p}$ are *different* tensors. The first gives the value $\tilde{p}(\vec{A})\tilde{q}(\vec{B})$, the second the value $\tilde{q}(\vec{A})\ \tilde{p}(\vec{B})$.

Components

The most general $\binom{0}{2}$ tensor is not a simple outer product, but it can always be represented as a sum of such tensors. To see this we must first consider the components of an arbitrary $\binom{0}{2}$ tensor **f**:

$$f_{\alpha\beta} := \mathbf{f}(\vec{e}_\alpha, \vec{e}_\beta). \tag{3.21}$$

Since each index can have four values, there are 16 components, and they can be thought of as being arrayed in a matrix. The value of **f** on arbitrary vectors is

$$\begin{aligned} \mathbf{f}(\vec{A}, \vec{B}) &= \mathbf{f}(A^\alpha \vec{e}_\alpha, B^\beta \vec{e}_\beta) \\ &= A^\alpha B^\beta \mathbf{f}(\vec{e}_\alpha, \vec{e}_\beta) \\ &= A^\alpha B^\beta f_{\alpha\beta}. \end{aligned} \tag{3.22}$$

(Again notice that two different dummy indices are used to keep the different summations distinct.) Can we form a basis for these tensors? That is, can we define a set of 16 $\binom{0}{2}$ tensors $\tilde{\omega}^{\alpha\beta}$ such that, analogous to Eq. (3.11),

$$\mathbf{f} = f_{\alpha\beta}\tilde{\omega}^{\alpha\beta}? \tag{3.23}$$

For this to be the case we would have to have

$$f_{\mu\nu} = \mathbf{f}(\vec{e}_\mu, \vec{e}_\nu) = f_{\alpha\beta}\tilde{\omega}^{\alpha\beta}(\vec{e}_\mu, \vec{e}_\nu)$$

and this would imply, as before, that

$$\tilde{\omega}^{\alpha\beta}(\vec{e}_\mu, \vec{e}_\nu) = \delta^\alpha{}_\mu \delta^\beta{}_\nu. \tag{3.24}$$

But $\delta^\alpha{}_\mu$ is (by Eq. (3.12)) the value of $\tilde{\omega}^\alpha$ on $\vec{e}_\mu$, and analogously for $\delta^\beta{}_\nu$. Therefore, $\tilde{\omega}^{\alpha\beta}$ is a tensor the value of which is just the product of the values of two basis one-forms, and we therefore conclude

$$\tilde{\omega}^{\alpha\beta} = \tilde{\omega}^\alpha \otimes \tilde{\omega}^\beta. \tag{3.25}$$

So the tensors $\tilde{\omega}^\alpha \otimes \tilde{\omega}^\beta$ are a basis for all $\binom{0}{2}$ tensors, and we can write

$$\mathbf{f} = f_{\alpha\beta}\tilde{\omega}^\alpha \otimes \tilde{\omega}^\beta. \tag{3.26}$$

This is one way in which a general $\binom{0}{2}$ tensor is a sum over simple outer-product tensors.

Symmetries

A $\binom{0}{2}$ tensor takes two arguments, and their order is important, as we have seen. The behavior of the value of a tensor under an interchange of its arguments is an important property of it. A tensor **f** is called *symmetric* if

$$\mathbf{f}(\vec{A}, \vec{B}) = \mathbf{f}(\vec{B}, \vec{A}) \quad \forall \vec{A}, \vec{B}. \tag{3.27}$$

Setting $\vec{A} = \vec{e}_\alpha$ and $\vec{B} = \vec{e}_\beta$, this implies of its components that

$$f_{\alpha\beta} = f_{\beta\alpha}. \tag{3.28}$$

This is the same as the condition that the matrix array of the elements is symmetric. An arbitrary $\binom{0}{2}$ tensor $\mathbf{h}$ can define a new symmetric $\mathbf{h}_{(s)}$ by the rule

$$\mathbf{h}_{(s)}(\vec{A}, \vec{B}) = \tfrac{1}{2}\mathbf{h}(\vec{A}, \vec{B}) + \tfrac{1}{2}\mathbf{h}(\vec{B}, \vec{A}). \tag{3.29}$$

Make sure you understand that $\mathbf{h}_{(s)}$ satisfies Eq. (3.27) above. For the components this implies

$$h_{(s)\alpha\beta} = \tfrac{1}{2}(h_{\alpha\beta} + h_{\beta\alpha}). \tag{3.30}$$

This is such an important mathematical property that a special notation is used for it:

$$h_{(\alpha\beta)} := \tfrac{1}{2}(h_{\alpha\beta} + h_{\beta\alpha}). \tag{3.31}$$

Therefore, the numbers $h_{(\alpha\beta)}$ are the components of the symmetric tensor formed from $\mathbf{h}$.

Similarly, a tensor $\mathbf{f}$ is called *antisymmetric* if

$$\mathbf{f}(\vec{A}, \vec{B}) = -\mathbf{f}(\vec{B}, \vec{A}), \quad \forall \vec{A}, \vec{B}, \tag{3.32}$$

$$f_{\alpha\beta} = -f_{\beta\alpha}. \tag{3.33}$$

An antisymmetric $\binom{0}{2}$ tensor can always be formed as

$$\mathbf{h}_{(A)}(\vec{A}, \vec{B}) = \tfrac{1}{2}\mathbf{h}(\vec{A}, \vec{B}) - \tfrac{1}{2}\mathbf{h}(\vec{B}, \vec{A}),$$
$$h_{(A)\alpha\beta} = \tfrac{1}{2}(h_{\alpha\beta} - h_{\beta\alpha}).$$

The notation here is to use square brackets on the indices:

$$h_{[\alpha\beta]} = \tfrac{1}{2}(h_{\alpha\beta} - h_{\beta\alpha}). \tag{3.34}$$

Notice that

$$\begin{aligned} h_{\alpha\beta} &= \tfrac{1}{2}(h_{\alpha\beta} + h_{\beta\alpha}) + \tfrac{1}{2}(h_{\alpha\beta} - h_{\beta\alpha}) \\ &= h_{(\alpha\beta)} + h_{[\alpha\beta]}. \end{aligned} \tag{3.35}$$

So any $\binom{0}{2}$ tensor can be split *uniquely* into its symmetric and antisymmetric parts.

The metric tensor **g** is symmetric, as can be deduced from Eq. (2.26):

$$\mathbf{g}(\vec{A},\vec{B}) = \mathbf{g}(\vec{B},\vec{A}). \tag{3.36}$$

3.5 Metric as a mapping of vectors into one-forms

We now introduce what we shall later see is the fundamental role of the metric in differential geometry, to act as a mapping between vectors and one-forms. To see how this works, consider **g** and a single vector $\vec{V}$. Since **g** requires two vectorial arguments, the expression $\mathbf{g}(\vec{V},\ )$ still lacks one: when another one is supplied, it becomes a number. Therefore, $\mathbf{g}(\vec{V},\ )$ considered as a function of vectors (which are to fill in the empty 'slot' in it) is a linear function of vectors producing real numbers: a one-form. We call it $\tilde{V}$:

$$\mathbf{g}(\vec{V},\quad) := \tilde{V}(\quad), \tag{3.37}$$

where blanks inside parentheses are a way of indicating that a vector argument is to be supplied. Then $\tilde{V}$ is the one-form that evaluates on a vector $\vec{A}$ to $\vec{V}\cdot\vec{A}$:

$$\tilde{V}(\vec{A}) := \mathbf{g}(\vec{V},\vec{A}) = \vec{V}\cdot\vec{A}. \tag{3.38}$$

Note that since **g** is symmetric, we also can write

$$\mathbf{g}(\quad,\vec{V}) := \tilde{V}(\quad).$$

What are the components of $\tilde{V}$? They are

$$\begin{aligned} V_\alpha := \tilde{V}(\vec{e}_\alpha) &= \vec{V}\cdot\vec{e}_\alpha = \vec{e}_\alpha\cdot\vec{V} \\ &= \vec{e}_\alpha\cdot(V^\beta\vec{e}_\beta) \\ &= (\vec{e}_\alpha\cdot\vec{e}_\beta)V^\beta \end{aligned}$$

$$V_\alpha = \eta_{\alpha\beta}V^\beta. \tag{3.39}$$

It is important to notice here that we distinguish the components V^α of $\vec{V}$ from the components V_β of $\tilde{V}$ *only* by the position of the index: on a vector it is up; on a one-form, down. Then, from Eq. (3.39), we have as a special case

$$\begin{aligned} V_0 = V^\beta \eta_{\beta 0} &= V^0 \eta_{00} + V^1 \eta_{10} + \dots \\ &= V^0(-1) + 0 + 0 + 0 \\ &= -V^0, \end{aligned} \tag{3.40}$$

$$\begin{aligned} V_1 = V^\beta \eta_{\beta 1} &= V^0 \eta_{01} + V^1 \eta_{11} + \dots \\ &= +V^1, \end{aligned} \tag{3.41}$$

etc. This may be summarized as:

$$\begin{aligned} \text{if } \vec{V} &\to (a, b, c, d), \\ \text{then } \tilde{V} &\to (-a, b, c, d). \end{aligned} \tag{3.42}$$

The components of $\tilde{V}$ are obtained from those of $\vec{V}$ by changing the sign of the time component. (Since this depended upon the components $\eta_{\alpha\beta}$, in situations we encounter later, where the metric has more complicated components, this rule of correspondence between $\tilde{V}$ and $\vec{V}$ will also be more complicated.) The inverse: going from $\tilde{\boldsymbol{A}}$ to $\vec{\boldsymbol{A}}$

Does the metric also provide a way of finding a vector $\vec{A}$ that is related to a given one-form $\tilde{A}$? The answer is yes. Consider Eq. (3.39). It says that $\{V_\alpha\}$ is obtained by multiplying $\{V^\beta\}$ by a matrix $(\eta_{\alpha\beta})$. If this matrix has an inverse, then we could use it to obtain $\{V^\beta\}$ from $\{V_\alpha\}$. This inverse exists if and only if $(\eta_{\alpha\beta})$ has nonvanishing determinant. But since $(\eta_{\alpha\beta})$ is a diagonal matrix with entries $(-1, 1, 1, 1)$, its determinant is simply -1. An inverse does exist, and we call its components $\eta^{\alpha\beta}$. Then, given $\{A_\beta\}$ we can find $\{A^\alpha\}$:

$$A^\alpha := \eta^{\alpha\beta} A_\beta. \tag{3.43}$$

The use of the inverse guarantees that the two sets of components satisfy Eq. (3.39): $A_\beta = \eta_{\beta\alpha} A^\alpha.$

So the mapping provided by **g** between vectors and one-forms is one-to-one and invertible.

In particular, with $\tilde{\mathrm{d}}\phi$ we can associate a vector $\vec{\mathrm{d}}\phi$, which is the one usually associated with the gradient. We can see that this vector is orthogonal to surfaces of constant ϕ as follows: its inner product with a vector in a surface of constant

ϕ is, by this mapping, identical with the value of the one-form $\tilde{d}\phi$ on that vector. This, in turn, must be zero since $\tilde{d}\phi(\vec{V})$ is the rate of change of ϕ along $\vec{V}$, which in this case is zero since $\vec{V}$ is taken to be in a surface of constant ϕ.

It is important to know what $\{\eta^{\alpha\beta}\}$ is. You can easily verify that

$$\eta^{00} = -1, \quad \eta^{0i} = 0, \quad \eta^{ij} = \delta^{ij}, \tag{3.44}$$

so that $(\eta^{\alpha\beta})$ is *identical* to $(\eta_{\alpha\beta})$. Thus, to go from a one-form to a vector, simply change the sign of the time component.

Why distinguish one-forms from vectors?

In Euclidean space, in Cartesian coordinates the metric is just $\{\delta_{ij}\}$, so the components of one-forms and vectors are the same. Therefore no distinction is ever made in elementary vector algebra. But in SR the components differ (by that one change in sign). Therefore, whereas the gradient has components

$$\tilde{d}\phi \to \left(\frac{\partial\phi}{\partial t}, \frac{\partial\phi}{\partial x}, \ldots\right),$$

the associated vector normal to surfaces of constant ϕ has components

$$\vec{d}\phi \to \left(-\frac{\partial\phi}{\partial t}, \frac{\partial\phi}{\partial x}, \ldots\right). \tag{3.45}$$

Had we simply tried to *define* the ‘vector gradient’ of a function as the vector with these components, without first discussing one-forms, the reader would have been justified in being more than a little skeptical. The non-Euclidean metric of SR forces us to be aware of the basic distinction between one-forms and vectors: it can’t be swept under the rug.

As we remarked earlier, vectors and one-forms are dual to one another. Such dual spaces are important and are found elsewhere in mathematical physics. The simplest example is the space of column vectors in matrix algebra

$$\begin{pmatrix} a \\ b \\ \vdots \end{pmatrix},$$

whose dual space is the space of row vectors $(a\ b\ \cdots)$. Notice that the product

$$(a\,b\,\ldots)\begin{pmatrix} p \\ q \\ \vdots \end{pmatrix} = ap + bq + \ldots \tag{3.46}$$

is a real number, so that a row vector can be considered to be a one-form on column vectors. The operation of finding an element of one space from one of the others is called the 'adjoint' and is 1–1 and invertible. A less trivial example arises in quantum mechanics. A wave-function (probability amplitude that is a solution to Schrödinger's equation) is a complex scalar field $\psi(\vec{x})$, and is drawn from the *Hilbert space* of all such functions. This Hilbert space is a vector space, since its elements (functions) satisfy the axioms of a vector space. What is the dual space of one-forms? The crucial hint is that the inner product of any two functions $\phi(\vec{x})$ and $\psi(\vec{x})$ is *not* $\int \phi(\vec{x})\psi(\vec{x})\, \mathrm{d}^3x$, but, rather, is $\int \phi^*(\vec{x})\psi(\vec{x})\, \mathrm{d}^3x$, the asterisk denoting complex conjugation. The function $\phi^*(\vec{x})$ acts like a one-form whose value on $\psi(\vec{x})$ is its integral with it (analogous to the sum in Eq. (3.8)). The operation of complex conjugation acts like our metric tensor, transforming a vector $\phi(\vec{x})$ (in the Hilbert space) into a one-form $\phi^*(\vec{x})$. The fact that $\phi^*(\vec{x})$ is also a function in the Hilbert space is, at this level, a distraction. (It is equivalent to saying that members of the set (1, −1, 0, 0) can be components of either a vector or a one-form.) The important point is that in the integral $\int \phi^*(\vec{x})\psi(\vec{x})\, \mathrm{d}^3x$, the function $\phi^*(\vec{x})$ is acting as a one-form, producing a (complex) number from the vector $\psi(\vec{x})$. This dualism is most clearly brought out in the Dirac 'bra' and 'ket' notation. Elements of the space of all states of the system are called $|\ \rangle$ (with identifying labels written inside), while the elements of the dual (adjoint with complex conjugate) space are called $\langle\ |$. Two 'vectors' $|1\rangle$ and $|2\rangle$ don't form a number, but a vector and a dual vector $|1\rangle$ and $\langle 2|$ do: $\langle 2|1\rangle$ is the name of this number.

In such ways the concept of a dual vector space arises very frequently in advanced mathematical physics.

Magnitudes and scalar products of one-forms

A one-form $\tilde{p}$ is defined to have the same magnitude as its associated vector $\vec{p}$. Thus we write

$$\tilde{p}^2 = \vec{p}^2 = \eta_{\alpha\beta}p^\alpha p^\beta. \tag{3.47}$$

This would seem to involve finding $\{p^\alpha\}$ from $\{p_\alpha\}$ before using Eq. (3.47), but we can easily get around this. We use Eq. (3.43) for both p^α and p^β in Eq. (3.47):

$$\tilde{p}^2 = \eta_{\alpha\beta}(\eta^{\alpha\mu}p_\mu)(\eta^{\beta\nu}p_\nu). \tag{3.48}$$

(Notice that each independent summation uses a different dummy index.) But since $\eta_{\alpha\beta}$ and $\eta^{\beta\nu}$ are inverse matrices to each other, the sum on β collapses:

$$\eta_{\alpha\beta}\eta^{\beta\nu} = \delta^\nu{}_\alpha. \tag{3.49}$$

Using this in Eq. (3.48) gives

$$\tilde{p}^2 = \eta^{\alpha\mu}p_\mu p_\alpha. \tag{3.50}$$

Thus, the inverse metric tensor can be used directly to find the magnitude of $\tilde{p}$ from its components. We can use Eq. (3.44) to write this explicitly as

$$\tilde{p}^2 = -(p_0)^2 + (p_1)^2 + (p_2)^2 + (p_3)^2. \tag{3.51}$$

This is the same rule, in fact, as Eq. (2.24) for vectors. By its definition, this is frame invariant. One-forms are timelike, spacelike, or null, as their associated vectors are.

As with vectors, we can now define an inner product of one-forms. This is

$$\tilde{p} \cdot \tilde{q} := \tfrac{1}{2}\left[(\tilde{p} + \tilde{q})^2 - \tilde{p}^2 - \tilde{q}^2\right]. \tag{3.52}$$

Its expression in terms of components is, not surprisingly,

$$\tilde{p} \cdot \tilde{q} = -p_0q_0 + p_1q_1 + p_2q_2 + p_3q_3. \tag{3.53}$$

Normal vectors and unit normal one-forms

A vector is said to be normal to a surface if its associated one-form is a normal one-form. Eq. (3.38) shows that this definition is equivalent to the usual one that the vector be orthogonal to all tangent vectors. A normal vector or one-form is said to be a *unit normal* if its magnitude is ±1. (We can't demand that it be +1, since timelike vectors will have negative magnitudes. All we can do is to multiply the vector or form by an overall factor to scale its magnitude to ±1.) Note that null normals cannot be unit normals.

A three-dimensional surface is said to be timelike, spacelike, or null according to which of these classes its normal falls into. (Exer. 12, § 3.10, proves that this definition is self-consistent.) In Exer. 21, § 3.10, we explore the following curious properties normal vectors have on account of our metric. An outward normal vector is the vector associated with an outward normal one-form, as defined earlier. This ensures that its scalar product with any vector which points outwards is positive. If the surface is spacelike, the outward normal vector points outwards. If the surface is timelike, however, the outward normal vector points *inwards*. And if the surface is null, the outward vector is *tangent* to the surface! These peculiarities simply reinforce the view that it is more natural to regard the normal as a one-form, where the metric doesn't enter the definition.

3.6 Finally: $\binom{M}{N}$ tensors

Vector as a function of one-forms

The dualism discussed above is in fact complete. Although we defined one-forms as functions of vectors, we can now see that vectors can perfectly well be regarded as linear functions that map one-forms into real numbers. Given a vector $\vec{V}$, once we supply a one-form we get a real number:

$$\vec{V}(\tilde{p}) \equiv \tilde{p}(\vec{V}) \equiv p_\alpha V^\alpha \equiv \langle \tilde{p}, \vec{V} \rangle. \tag{3.54}$$

In this way we dethrone vectors from their special position as things 'acted on' by tensors, and regard them as tensors themselves, specifically as linear functions of single one-forms into real numbers. The last notation on Eq. (3.54) is new, and emphasizes the equal status of the two objects.

$\binom{M}{0}$ tensors

Generalizing this, we define:

An $\binom{M}{0}$ tensor is a linear function of M one-forms into the real numbers.

All our previous discussions of $\binom{0}{N}$ tensors apply here. A simple $\binom{2}{0}$ tensor is $\vec{V} \otimes \vec{W}$, which, when supplied with two arguments $\tilde{p}$ and $\tilde{q}$, gives the number $\vec{V}(\tilde{p})\vec{W}(\tilde{q}) := \tilde{p}(\vec{V})\tilde{q}(\vec{W}) = V^\alpha p_\alpha W^\beta q_\beta$. So $\vec{V} \otimes \vec{W}$ has components $V^\alpha W^\beta$. A basis for $\binom{2}{0}$ tensors is $\vec{e}_\alpha \otimes \vec{e}_\beta$. The components of an $\binom{M}{0}$ tensor are its values when the basis one-form $\tilde{\omega}^\alpha$ are its arguments. Notice that $\binom{M}{0}$ tensors have components all of whose indices are superscripts.

$\binom{M}{N}$ tensors

The final generalization is:

An $\binom{M}{N}$ tensor is a linear function of *M* one-forms *and* *N* vectors into the real numbers.

For instance, if **R** is a $\binom{1}{1}$ tensor, then it requires a one-form $\tilde{p}$ and a vector $\vec{A}$ to give a number $\mathbf{R}(\tilde{p};\ \vec{A})$. It has components $\mathbf{R}(\tilde{\omega}^{\alpha};\ \vec{e}_{\beta}) := \boldsymbol{R}^{\alpha}{}_{\beta}$. In general, the components of a $\binom{M}{N}$ tensor will have *M* indices up and *N* down. In a new frame,

$$\begin{aligned} R^{\bar{\alpha}}{}_{\bar{\beta}} &= \mathbf{R}(\tilde{\omega}^{\bar{\alpha}}; \vec{e}_{\bar{\beta}}) \\ &= \mathbf{R}(\Lambda^{\bar{\alpha}}{}_{\mu}\tilde{\omega}^{\mu}; \Lambda^{\nu}{}_{\bar{\beta}}\vec{e}_{\nu}) \\ &= \Lambda^{\bar{\alpha}}{}_{\mu}\Lambda^{\nu}{}_{\bar{\beta}}\mathbf{R}^{\mu}{}_{\nu}. \end{aligned} \tag{3.55}$$

So the transformation of components is simple: each index transforms by bringing in a Λ whose indices are arranged in the only way permitted by the summation convention. Some old names that are still in current use are: upper indices are called 'contravariant' (because they transform *contrary* to basis vectors) and lower ones 'covariant'. An $\binom{M}{N}$ tensor is said *N* to be '*M*-times contravariant and *N*-times covariant'.

Circular reasoning?

At this point the student might worry that all of tensor algebra has become circular: one-forms were defined in terms of vectors, but now we have defined vectors in terms of one-forms. This 'duality' is at the heart of the theory, but is not circularity. It means we can do as physicists do, which is to identify the vectors with displacements $\Delta\vec{x}$ and things like it (such as $\vec{p}$ and $\vec{v}$ and then generate all $\binom{M}{N}$ tensors by the rules of tensor algebra; these tensors inherit a physical meaning from the original meaning we gave vectors. But we could equally well have associated one-forms with some physical objects (gradients, for example) and recovered the whole algebra from that starting point. The power of the mathematics is that it doesn't need (or want) to say *what* the original vectors or one-forms are. It simply gives rules for manipulating them. The association of, say, $\vec{p}$ with a vector is at the interface between physics and mathematics: it is how we make a mathematical model of the physical world. A geometer does the same. He adds to the notion of these abstract tensor spaces the idea of what a vector in a curved space is. The modern geometer's idea of a vector is something we shall learn about when we come to curved spaces. For now we will get some practice with tensors in physical situations, where we stick with our (admittedly imprecise) notion of vectors 'like' $\Delta\vec{x}$.

3.7 Index 'raising' and 'lowering'

In the same way that the metric maps a vector $\vec{V}$ into a one-form $\tilde{V}$, it maps an $\binom{N}{M}$ tensor into an $\binom{N-1}{M+1}$ tensor. Similarly, the inverse maps an $\binom{N}{M}$ tensor into an $\binom{N+1}{M-1}$ tensor. Normally, these are given the same name, and are distinguished only by the positions of their indices. Suppose $T^{\alpha\beta}{}_{\gamma}$ are the components of a $\binom{2}{1}$ tensor. Then

$$T^{\alpha}{}_{\beta\gamma} := \eta_{\beta\mu} T^{\alpha\mu}{}_{\gamma} \tag{3.56}$$

are the components of a $\binom{1}{2}$ tensor (obtained by mapping the *second* one-form argument of $T^{\alpha\beta}{}_{\gamma}$ into a vector), and

$$T_{\alpha}{}^{\beta}{}_{\gamma} := \eta_{\alpha\mu} T^{\mu\beta}{}_{\gamma} \tag{3.57}$$

are the components of another (inequivalent) $\binom{1}{2}$ tensor (mapping on the *first* index), while

$$T^{\alpha\beta\gamma} := \eta^{\gamma\mu} T^{\alpha\beta}{}_{\mu} \tag{3.58}$$

are the components of a $\binom{3}{0}$ tensor. These operations are, naturally enough, called index 'raising' and 'lowering'. Whenever we speak of raising or lowering an index we mean this map generated by the metric. The rule in SR is simple: when raising or lowering a '0' index, the sign of the component changes; when raising or lowering a '1' or '2' or '3' index (in general, an 'i' index) the component is unchanged.

Mixed components of metric

The numbers $\{\eta_{\alpha\beta}\}$ are the components of the metric, and $\{\eta^{\alpha\beta}\}$ those of its inverse. Suppose we raise an index of $\eta_{\alpha\beta}$ using the inverse. Then we get the 'mixed' components of the metric,

$$\eta^{\alpha}{}_{\beta} \equiv \eta^{\alpha\mu}\eta_{\mu\beta}. \tag{3.59}$$

But on the right we have just the matrix product of two matrices that are the inverse of each other (readers who aren't sure of this should verify the following equation by direct calculation), so it is the unit identity matrix. Since one index is up and one down, it is the Kronecker delta, written as

$$\eta^{\alpha}{}_{\beta} \equiv \delta^{\alpha}{}_{\beta}. \tag{3.60}$$

By raising the other index we merely obtain an identity, $\eta^{\alpha\beta} = \eta^{\alpha\beta}$. So we can regard $\eta^{\alpha\beta}$ as the components of the $\binom{2}{0}$ tensor, which is mapped from the $\binom{0}{2}$ tensor $\mathbf{g}$ by $\mathbf{g}^{-1}$. So, for $\mathbf{g}$, its 'contravariant' components equal the elements of the matrix inverse of the matrix of its 'covariant' components. It is the only tensor for which this is true.

Metric and nonmetric vector algebras

It is of some interest to ask why the metric is the one that generates the correspondence between one-forms and vectors. Why not some other $\binom{0}{2}$ tensor that has an inverse? We'll explore that idea in stages.

First, why a correspondence at all? Suppose we had a 'nonmetric' vector algebra, complete with all the dual spaces and $\binom{M}{N}$ tensors. Why make a correspondence between one-forms and vectors? The answer is that sometimes we do and sometimes we don't. Without one, the inner product of two vectors is undefined, since numbers are produced only when one-forms act on vectors and vice-versa. In physics, scalar products are useful, so we need a metric. But there are *some* vector spaces in mathematical physics where metrics are not important. An example is phase space of classical and quantum mechanics.

Second, why the metric and not another tensor? If a metric were not defined but another symmetric tensor did the mapping, a mathematician would just call the other tensor the metric. That is, he would define it as the one generating a mapping. To a mathematician, the metric is an added bit of *structure* in the vector algebra. Different spaces in mathematics can have different metric structures. A *Riemannian* space is characterized by a metric that gives positive-definite magnitudes of vectors. One like ours, with indefinite sign, is called *pseudo-Riemannian*. We can even define a 'metric' that is *antisymmetric*: a two-dimensional space called *spinor space* has such a metric, and it turns out to be of fundamental importance in physics. But its structure is outside the scope of this book. The point here is that we don't have SR if we just discuss vectors and tensors. We get SR when we say that we have a metric with components $\eta_{\alpha\beta}$. If we assigned other components, we might get other spaces, in particular the curved spacetime of GR.

3.8 Differentiation of tensors

A function f is a $\binom{0}{0}$ tensor, and its gradient $\tilde{\mathrm{d}}f$ is a $\binom{0}{1}$ tensor. Differentiation of a function produces a tensor of one higher (covariant) rank. We shall now see that this applies as well to differentiation of tensors of *any* rank.

Consider a $\binom{1}{1}$ tensor $\mathbf{T}$ whose components $\{T^\alpha{}_\beta\}$ are functions of position. We can write $\mathbf{T}$ as

$$\mathbf{T} = T^\alpha{}_\beta \tilde{\omega}^\beta \otimes \vec{e}_\alpha. \tag{3.61}$$

Suppose, as we did for functions, that we move along a world line with parameter τ, proper time. The rate of change of $\mathbf{T}$,

$$\frac{\mathrm{d}\mathbf{T}}{\mathrm{d}\tau} = \lim_{\Delta\tau \to 0} \frac{\mathbf{T}(\tau + \Delta\tau) - \mathbf{T}(\tau)}{\Delta\tau}, \tag{3.62}$$

is not hard to calculate. Since the basis one-forms and vectors are the same everywhere (i.e. $\tilde{\omega}^\alpha(\tau + \Delta\tau) = \tilde{\omega}^\alpha(\tau)$), it follows that

$$\frac{\mathrm{d}\mathbf{T}}{\mathrm{d}\tau} = \left(\frac{\mathrm{d}T^\alpha{}_\beta}{\mathrm{d}\tau}\right) \tilde{\omega}^\beta \otimes \vec{e}_\alpha, \tag{3.63}$$

where $\mathrm{d}T^\alpha{}_\beta/\mathrm{d}\tau$ is the ordinary derivative of the function $T^\alpha{}_\beta$ along the world line:

$$\mathrm{d}T^\alpha{}_\beta/\mathrm{d}\tau = T^\alpha{}_{\beta,\gamma}\, U^\gamma. \tag{3.64}$$

Now, the object $\mathrm{d}\mathbf{T}/\mathrm{d}\tau$ is a $\binom{1}{1}$ tensor, since in Eq. (3.62) it is defined to be just the difference between two such tensors. From Eqs. (3.63) and (3.64) we have, for any vector $\vec{U}$,

$$\mathrm{d}\mathbf{T}/\mathrm{d}\tau = (T^\alpha{}_{\beta,\gamma} \tilde{\omega}^\beta \otimes \vec{e}_\alpha)\, U^\gamma, \tag{3.65}$$

from which we can deduce that

$$\nabla\mathbf{T} := (T^\alpha{}_{\beta,\gamma} \tilde{\omega}^\beta \otimes \tilde{\omega}^\gamma \otimes \vec{e}_\alpha) \tag{3.66}$$

is a $\binom{1}{2}$ tensor. This tensor is called the gradient of $\mathbf{T}$.

We use the notation $\nabla\mathbf{T}$ rather than $\tilde{\mathrm{d}}\mathbf{T}$ because the latter notation is usually reserved by mathematicians for something else. We also have a convenient notation for Eq. (3.65):

$$\mathrm{d}\mathbf{T}/\mathrm{d}\tau = \nabla_{\vec{U}}\mathbf{T}, \tag{3.67}$$

$$\nabla_{\vec{U}}\mathbf{T} \to \left\{T^{\alpha}{}_{\beta,\gamma}U^{\gamma}\right\}. \tag{3.68}$$

This derivation made use of the fact that the basis vectors (and therefore the basis one-forms) were constant everywhere. We will find that we can't assume this in the curved spacetime of GR, and taking this into account will be our entry point into the theory !

3.9 Further reading

Our approach to tensor analysis stresses the geometrical nature of tensors rather than the transformation properties of their components. Students who wish amplification of some of the points here can consult the early chapters of Misner *et al.* (1973) or Schutz (1980b). See also Bishop and Goldberg (1981).

Most introductions to tensors for physicists outside relativity confine themselves to 'Cartesian' tensors, *i.e.* to tensor components in three-dimensional Cartesian coordinates. See, for example, Bourne and Kendall (1991) or the chapter in Mathews and Walker (1965).

A very complete reference for tensor analysis in the older style based upon coordinate transformations is Schouten (1990). See also Yano (1955). Books which develop that point of view for tensors in relativity include Adler *et al.* (1975), Landau and Lifshitz (1962), and Stephani (2004).

3.10 Exercises

a) Given an arbitrary set of numbers $\{M_{\alpha\beta};\ \alpha = 0, \ldots, 3;\ \beta = 0, \ldots, 3\}$ and two arbitrary sets of vector components $\{A^{\mu},\ \mu = 0, \ldots, 3\}$ and $\{B^{\nu},\ \nu = 0, \ldots, 3\}$, show that the two expressions

$$M_{\alpha\beta}A^{\alpha}B^{\beta} := \sum_{\alpha=0}^{3}\sum_{\beta=0}^{3} M_{\alpha\beta}A^{\alpha}B^{\beta}$$

and

$$\sum_{\alpha=0}^{3} M_{\alpha\alpha}A^{\alpha}B^{\alpha}$$

are not equivalent.

(b) Show that

$$A^\alpha B^\beta \eta_{\alpha\beta} = -A^0B^0 + A^1B^1 + A^2B^2 + A^3B^3.$$

Prove that the set of all one-forms is a vector space.

a) Prove, by writing out all the terms, the validity of the following

$$\tilde{p}(A^\alpha \vec{e}_\alpha) = A^\alpha \tilde{p}(\vec{e}_\alpha).$$

(b) Let the components of $\tilde{p}$ be (−1, 1, 2, 0), those of $\vec{A}$ be (2, 1, 0, −1) and those of $\vec{B}$ be (0, 2, 0, 0). Find (i) $\tilde{p}(\vec{A})$; (ii) $\tilde{p}(\vec{B})$; (iii) $\tilde{p}(\vec{A} - 3\vec{B})$; (iv) $\tilde{p}(\vec{A}) - 3\tilde{p}(\vec{B})$.

Given the following vectors in $\mathcal{O}$:

$$\vec{A} \underset{\mathcal{O}}{\to} (2, 1, 1, 0), \vec{B} \underset{\mathcal{O}}{\to} (1, 2, 0, 0), \vec{C} \underset{\mathcal{O}}{\to} (0, 0, 1, 1), \vec{D} \underset{\mathcal{O}}{\to} (-3, 2, 0, 0),$$

(a) show that they are linearly independent;

(b) find the components of $\tilde{p}$ if

$$\tilde{p}(\vec{A}) = 1, \tilde{p}(\vec{B}) = -1, \tilde{p}(\vec{C}) = -1, \tilde{p}(\vec{D}) = 0;$$

(c) find the value of $\tilde{p}(\vec{E})$ for $\vec{E} \underset{\mathcal{O}}{\to} (1, 1, 0, 0);$

(d) determine whether the one-forms $\tilde{p}$, $\tilde{q}$, $\tilde{r}$, and $\tilde{s}$ are linearly independent if $\tilde{q}(\vec{A}) = \tilde{q}(\vec{B}) = 0$, $\tilde{q}(\vec{C}) = 1$, $\tilde{q}(\vec{D}) = -1$, $\tilde{r}(\vec{A}) = 2$, $\tilde{r}(\vec{B}) = \tilde{r}(\vec{C}) = \tilde{r}(\vec{D}) = 0$, $\tilde{s}(\vec{A}) = -1$, $\tilde{s}(\vec{B}) = -1$, $\tilde{s}(\vec{C}) = \tilde{s}(\vec{D}) = 0$.

Justify each step leading from Eqs. (3.10a) to (3.10d).

Consider the basis $\{\vec{e}_\alpha\}$ of a frame $\mathcal{O}$ and the basis $(\tilde{\lambda}^0, \tilde{\lambda}^1, \tilde{\lambda}^2, \tilde{\lambda}^3)$ for the space of one-forms, where we have

$$\begin{aligned} \tilde{\lambda}^0 &\underset{\mathcal{O}}{\to} (1, 1, 0, 0), \\ \tilde{\lambda}^1 &\underset{\mathcal{O}}{\to} (1, -1, 0, 0), \\ \tilde{\lambda}^2 &\underset{\mathcal{O}}{\to} (0, 0, 1, -1), \\ \tilde{\lambda}^3 &\underset{\mathcal{O}}{\to} (0, 0, 1, 1). \end{aligned}$$

Note that $\{\tilde{\lambda}^\beta\}$ is *not* the basis dual to $\{\vec{e}_\alpha\}$.

(a) Show that $\tilde{p} \neq \tilde{p}(\vec{e}_\alpha)\tilde{\lambda}^\alpha$ for arbitrary $\tilde{p}$.

(b) Let $\tilde{p} \to_{\mathcal{O}} (1, 1, 1, 1)$. Find numbers l_α such that $\tilde{p} = l_\alpha \tilde{\lambda}^\alpha$.

These are the components of $\tilde{p}$ on $\{\tilde{\lambda}^\alpha\}$, which is to say that they are the values of $\tilde{p}$ on the elements of the vector basis dual to $\{\tilde{\lambda}^\alpha\}$.

Prove Eq. (3.13).

Draw the basis one-forms $\tilde{d}t$ and $\tilde{d}x$ of a frame $\mathcal{O}$.

Fig. 3.5 shows curves of equal temperature T (isotherms) of a metal plate. At the points $\mathcal{P}$ and $\mathcal{Q}$ as shown, estimate the components of the gradient $\tilde{d}T$. (Hint: the components are the contractions with the basis vectors, which can be estimated by counting the number of isotherms crossed by the vectors.)

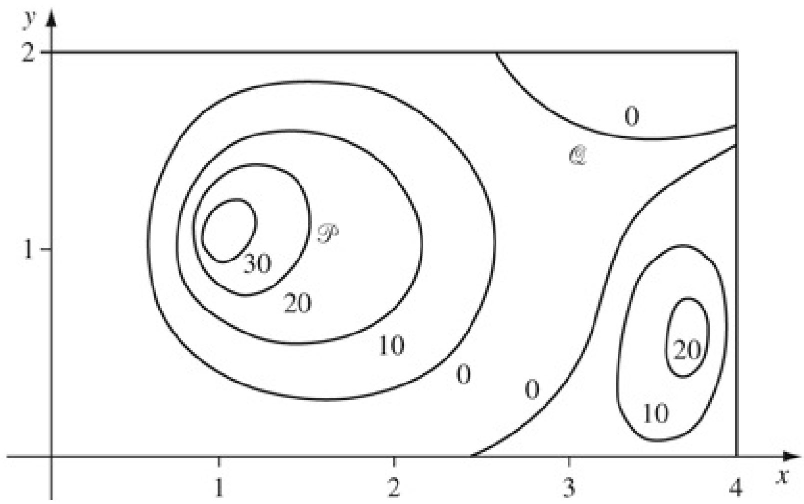


Figure 3.5 Isotherms of an irregularly heated plate.

) Given a frame $\mathcal{O}$ whose coordinates are $\{x^\alpha\}$, show that $\partial x^\alpha / \partial x^\beta = \delta^\alpha{}_\beta$.

(b) For any two frames, we have, Eq. (3.18):

$$\partial x^\beta / \partial x^{\bar{\alpha}} = \Lambda^\beta{}_{\bar{\alpha}}.$$

Show that (a) and the chain rule imply

$$\Lambda^\beta{}_{\bar{\alpha}} \Lambda^{\bar{\alpha}}{}_\mu = \delta^\beta{}_\mu.$$

This is the inverse property again.

Use the notation $\partial\phi/\partial x^\alpha = \phi_{,\alpha}$ to rewrite Eqs. (3.14), (3.15), and (3.18).

Let S be the two-dimensional plane $x = 0$ in three-dimensional Euclidean space. Let $\tilde{n} \neq 0$ be a normal one-form to S.

(a) Show that if $\vec{V}$ is a vector which is not tangent to S, then $\tilde{n}(\vec{V}) \neq 0$.

(b) Show that if $\tilde{n}(\vec{V}) > 0$, then $\tilde{n}(\vec{W}) > 0$ for any $\vec{W}$, which points toward the same side of S as $\vec{V}$ does (i.e. any $\vec{W}$ whose x components has the same sign as V^x).

(c) Show that any normal to S is a multiple of $\tilde{n}$.

(d) Generalize these statements to an arbitrary three-dimensional surface in four-dimensional spacetime.

Prove, by geometric or algebraic arguments, that $\tilde{d}f$ is normal to surfaces of constant f.

Let $\tilde{p} \to_{\mathcal{O}} (1, 1, 0, 0)$ and $\tilde{q} \to_{\mathcal{O}} (-1, 0, 1, 0)$ be two one-forms. Prove, by trying two vectors $\vec{A}$ and $\vec{B}$ as arguments, that $\tilde{p} \otimes \tilde{q} \neq \tilde{q} \otimes \tilde{p}$. Then find the components of $\tilde{p} \otimes \tilde{q}$.

Supply the reasoning leading from Eq. (3.23) to Eq. (3.24).

) Prove that $\mathbf{h}_{(s)}$ defined by

$$\mathbf{h}_{(s)}(\vec{A}, \vec{B}) = \tfrac{1}{2}\mathbf{h}(\vec{A}, \vec{B}) + \tfrac{1}{2}\mathbf{h}(\vec{B}, \vec{A}) \tag{3.69}$$

is an symmetric tensor.

(b) Prove that $\mathbf{h}_{(A)}$ defined by

$$\mathbf{h}_{(A)}(\vec{A}, \vec{B}) = \tfrac{1}{2}\mathbf{h}(\vec{A}, \vec{B}) - \tfrac{1}{2}\mathbf{h}(\vec{B}, \vec{A}) \tag{3.70}$$

is an antisymmetric tensor.

(c) Find the components of the symmetric and antisymmetric parts of $\tilde{p} \otimes \tilde{q}$ defined in Exer. 14.

(d) Prove that if $\mathbf{h}$ is an antisymmetric $\binom{0}{2}$ tensor,

$$\mathbf{h}(\vec{A}, \vec{A}) = 0$$

for any vector $\vec{A}$.

(e) Find the number of independent components $\mathbf{h}_{(s)}$ and $\mathbf{h}_{(A)}$ have.

) Suppose that $\mathbf{h}$ is a $\binom{0}{2}$ tensor with the property that, for *any* two vectors $\vec{A}$ and $\vec{B}$ (where $\vec{B} \neq 0$) $\mathbf{h}(\quad, \vec{A}) = \alpha \mathbf{h}(\quad, \vec{B})$,

where α is a number which may depend on $\vec{A}$ and $\vec{B}$. Show that there exist one-forms $\tilde{p}$ and $\tilde{q}$ such that $\mathbf{h} = \tilde{p} \otimes \tilde{q}$.

(b) Suppose $\mathbf{T}$ is a $\binom{1}{1}$ tensor, $\tilde{\omega}$ a one-form, $\vec{v}$ a vector, and $\mathbf{T}(\tilde{\omega}; \vec{v})$ the value of $\mathbf{T}$ on $\tilde{\omega}$ and $\vec{v}$. Prove that $\mathbf{T}(\quad; \vec{v})$ is a vector and $\mathbf{T}(\tilde{\omega}; \quad)$ is a one-form, *i.e.* that a $\binom{1}{1}$ tensor provides a map of vectors to vectors and one-forms to one-forms.

) Find the one-forms mapped by the metric tensor from the vectors

$$\vec{A} \to_{\mathcal{O}} (1, 0, -1, 0), \qquad \vec{B} \to_{\mathcal{O}} (0, 1, 1, 0), \qquad \vec{C} \to_{\mathcal{O}} (-1, 0, -1, 0),$$
$$\vec{D} \to_{\mathcal{O}} (0, 0, 1, 1).$$

(b) Find the vectors mapped by the inverse of the metric tensor from the one-form $\tilde{p} \to_{\mathcal{O}} (3, 0, -1, -1)$, $\tilde{q} \to_{\mathcal{O}} (1, -1, 1, 1)$, $\tilde{r} \to_{\mathcal{O}} (0, -5, -1, 0)$, $\tilde{s} \to_{\mathcal{O}} (-2, 1, 0, 0)$.

a) Prove that the matrix $\{\eta^{\alpha\beta}\}$ is inverse to $\{\eta_{\alpha\beta}\}$ by performing the matrix multiplication.

(b) Derive Eq. (3.53).

In Euclidean three-space in Cartesian coordinates, we don't normally distinguish between vectors and one-forms, because their components transform identically. Prove this in two steps.

(a) Show that

$$A^{\bar{\alpha}} = \Lambda^{\bar{\alpha}}{}_{\beta} A^{\beta}$$

and

$$P_{\bar{\beta}} = \Lambda^{\alpha}{}_{\bar{\beta}} P_{\alpha}$$

are the same transformation if the matrix $\{\Lambda^{\bar{\alpha}}{}_{\beta}\}$ equals the transpose of its inverse. Such a matrix is said to be *orthogonal*.

(b) The metric of such a space has components $\{\delta_{ij}, i, j = 1, \ldots, 3\}$. Prove that a transformation from one Cartesian coordinate system to another must obey

$$\delta_{\bar{i}\bar{j}} = \Lambda^{k}{}_{\bar{i}} \Lambda^{l}{}_{\bar{j}} \delta_{kl}$$

and that this implies $\{\Lambda^{k}{}_{\bar{i}}\}$ is an orthogonal matrix. See Exer. 32 for the analog of this in SR.

) Let a region of the $t - x$ plane be bounded by the lines $t = 0$, $t = 1$, $x = 0$, $x = 1$. Within the $t - x$ plane, find the unit outward normal one-forms and their associated vectors for each of the boundary lines.

(b) Let another region be bounded by the straight lines joining the events whose coordinates are (1, 0), (1, 1), and (2, 1). Find an outward normal for the null boundary and find its associated vector.

Suppose that instead of defining vectors first, we had begun by defining one-forms, aided by pictures like Fig. 3.4. Then we could have introduced vectors as linear real-valued functions of one-forms, and defined vector algebra by the analogs of Eqs. (3.6a) and (3.6b) (i.e. by exchanging arrows for tildes). Prove that, so defined, vectors form a vector space. This is another example of the duality between vectors and one-forms.

ı) Prove that the set of all $\binom{M}{N}$ tensors for fixed M, N forms a vector space. (You must N define addition of such tensors and their multiplication by numbers.) (b) Prove that a basis for this space is the set

$$\{\underbrace{\vec{e}_\alpha \otimes \vec{e}_\beta \otimes \cdots \otimes \vec{e}_\gamma}_{M \text{ vectors}} \otimes \underbrace{\tilde{\omega}^\mu \otimes \tilde{\omega}^\nu \otimes \cdots \otimes \tilde{\omega}^\lambda}_{N \text{ one-forms}}\}.$$

(You will have to define the outer product of more than two one-forms.)

) Given the components of a $\binom{2}{0}$ tensor $M^{\alpha\beta}$ as the matrix $\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & -1 & 0 & 2 \\ 2 & 0 & 0 & 1 \\ 1 & 0 & -2 & 0 \end{pmatrix}$,

find:

(i) the components of the symmetric tensor $M^{(\alpha\beta)}$ and the antisymmetric tensor $M^{[\alpha\beta]}$; (ii) the components of $M^\alpha{}_\beta$;

(iii) the components of $M_\alpha{}^\beta$; (iv) the components of $M_{\alpha\beta}$.

(b) For the $\binom{1}{1}$ tensor whose components are $M^\alpha{}_\beta$, does it make sense to speak of its symmetric and antisymmetric parts? If so, define them. If not, say why.

(c) Raise an index of the metric tensor to prove

$$\eta^\alpha{}_\beta = \delta^\alpha{}_\beta.$$

Show that if **A** is a $\binom{2}{0}$ tensor and **B** a $\binom{0}{2}$ tensor, then $A^{\alpha\beta}B_{\alpha\beta}$

is frame invariant, *i.e.* a scalar.

Suppose **A** is an antisymmetric $\binom{2}{0}$ tensor, **B** a symmetric $\binom{0}{2}$ tensor, **C** an arbitrary $\binom{0}{2}$ tensor, and **D** an arbitrary $\binom{2}{0}$ tensor. Prove: (a) $A^{\alpha\beta}B_{\alpha\beta} = 0$; (b) $A^{\alpha\beta}C_{\alpha\beta} = A^{\alpha\beta}C_{[\alpha\beta]}$; (c) $B_{\alpha\beta}D^{\alpha\beta} = B_{\alpha\beta}D^{(\alpha\beta)}$.

a) Suppose **A** is an antisymmetric $\binom{2}{0}$ tensor. Show that $\{A_{\alpha\beta}\}$, obtained by lowering indices by using the metric tensor, are components of an antisymmetric $\binom{0}{2}$ tensor.

(b) Suppose $V^\alpha = W^\alpha$. Prove that $V_\alpha = W_\alpha$.

Deduce Eq. (3.66) from Eq. (3.65).

Prove that tensor differentiation obeys the Leibniz (product) rule:

$$\nabla(\mathbf{A} \otimes \mathbf{B}) = (\nabla \mathbf{A}) \otimes \mathbf{B} + \mathbf{A} \otimes \nabla \mathbf{B}.$$

In some frame $\mathcal{O}$, the vector fields $\vec{U}$ and $\vec{D}$ have the components:
$\vec{U} \rightarrow (1 + t^2, t^2, \sqrt{2}\,t, 0)$,
$\vec{D} \rightarrow (x, 5\,tx, \sqrt{2}\,t, 0)$,

and the scalar ρ has the value

$$\rho = x^2 + t^2 - y^2.$$

(a) Find $\vec{U} \cdot \vec{U}$, $\vec{U} \cdot \vec{D}$, $\vec{D} \cdot \vec{D}$. Is $\vec{U}$ suitable as a four-velocity field? Is $\vec{D}$?

(b) Find the spatial velocity υ of a particle whose four-velocity is $\vec{U}$, for arbitrary t. What happens to it in the limits $t \to 0$, $t \to \infty$?

(c) Find U_α for all α.

(d) Find $U^\alpha{}_{,\beta}$ for all α, β.

(e) Show that $U_\alpha U^\alpha{}_{,\beta} = 0$ for all β. Show that $U^\alpha U_{\alpha,\beta} = 0$ for all β.

(f) Find $D^\beta{}_{,\beta}$.

(g) Find $(U^\alpha D^\beta)_{,\beta}$ for all α.

(h) Find $U_\alpha(U^\alpha D^\beta)_{,\beta}$ and compare with (f) above. Why are the two answers similar?

(i) Find $\rho_{,\alpha}$ for all α. Find $\rho^{,\alpha}$ for all α. (Recall that $\rho^{,\alpha} := \eta^{\alpha\beta}\rho_{,\beta}$.) What are the numbers $\{\rho^{,\alpha}\}$ the components of?

(j) Find $\nabla_{\vec{U}}\rho, \nabla_{\vec{U}}\vec{D}, \nabla_{\vec{D}}\rho, \nabla_{\vec{D}}\vec{U}$.

Consider a timelike unit four-vector $\vec{U}$, and the tensor **P** whose components are given by $P_{\mu\nu} = \eta_{\mu\nu} + U_\mu U_\nu$.

(a) Show that **P** is a projection operator that projects an arbitrary vector $\vec{V}$ into one orthogonal to $\vec{U}$. That is, show that the vector $\vec{V}_\perp$ whose components are
$V^\alpha_\perp = P^\alpha{}_\beta V^\beta = (\eta^\alpha{}_\beta + U^\alpha U_\beta)V^\beta$

is

(i) orthogonal to $\vec{U}$,

and

(ii) unaffected by **P**:

$$V_{\perp\perp}^{\alpha} := P^{\alpha}{}_{\beta} V_{\perp}^{\beta} = V_{\perp}^{\alpha}.$$

(b) Show that for an arbitrary non-null vector $\vec{q}$, the tensor that projects orthogonally to it has components $\eta_{\mu\nu} - q_{\mu}q_{\nu}/(q^{\alpha}q_{\alpha})$.

How does this fail for null vectors? How does this relate to the definition of **P**?

(c) Show that **P** defined above is the metric tensor for vectors perpendicular to $\vec{U}$:

$$\begin{aligned}\mathbf{P}(\vec{V}_{\perp}, \vec{W}_{\perp}) &= \mathbf{g}(\vec{V}_{\perp}, \vec{W}_{\perp}) \\ &= \vec{V}_{\perp} \cdot \vec{W}_{\perp}.\end{aligned}$$

) From the definition $f_{\alpha\beta} = \mathbf{f}(\vec{e}_{\alpha}, \vec{e}_{\beta})$ for the components of a $\binom{0}{2}$ tensor, prove that the transformation law is $f_{\bar{\alpha}\bar{\beta}} = \Lambda^{\mu}{}_{\bar{\alpha}} \Lambda^{\nu}{}_{\bar{\beta}} f_{\mu\nu}$

and that the matrix version of this is

$$(\bar{f}) = (\Lambda)^{T}(f)(\Lambda),$$

where (Λ) is the matrix with components $\Lambda^{\mu}{}_{\bar{\alpha}}$.

(b) Since our definition of a Lorentz frame led us to deduce that the metric tensor has components $\eta_{\alpha\beta}$, this must be true in all Lorentz frames. We are thus led to a more general *definition* of a Lorentz transformation as one whose matrix $\Lambda^{\mu}{}_{\bar{\alpha}}$ satisfies

$$\eta_{\bar{\alpha}\bar{\beta}} = \Lambda^{\mu}{}_{\bar{\alpha}} \Lambda^{\nu}{}_{\bar{\beta}} \eta_{\mu\nu}. \tag{3.71}$$

Prove that the matrix for a boost of velocity $\upsilon\, \vec{e}_{x}$ satisfies this, so that this new definition includes our older one.

(c) Suppose (Λ) and (L) are two matrices which satisfy Eq. (3.71), *i.e.* $(\eta) = (\Lambda)^{T}(\eta)(\Lambda)$ and similarly for (L). Prove that $(\Lambda)(L)$ is also the matrix of a Lorentz transformation.

The result of Exer. 32c establishes that Lorentz transformations form a group, represented by multiplication of their matrices. This is called the *Lorentz group*, denoted by $L(4)$ or 0(1,3).

(a) Find the matrices of the identity element of the Lorentz group and of the element inverse to that whose matrix is implicit in Eq. (1.12).

(b) Prove that the determinant of any matrix representing a Lorentz transformation is ±1.

(c) Prove that those elements whose matrices have determinant +1 form a subgroup, while those with −1 do not.

(d) The three-dimensional orthogonal group $O(3)$ is the analogous group for the metric of three-dimensional Euclidean space. In Exer. 20b, we saw that it was represented by the orthogonal matrices. Show that the orthogonal matrices do form a group, and then show that 0(3) is (isomorphic to) a subgroup of $L(4)$.

Consider the coordinates $u = t - x$, $v = t + x$ in Minkowski space.

(a) Define $\vec{e}_u$ to be the vector connecting the events with coordinates $\{u = 1, v = 0, y = 0, z = 0\}$ and $\{u = 0, v = 0, y = 0, z = 0\}$, and analogously for $\vec{e}_v$. Show that $\vec{e}_u = (\vec{e}_t - \vec{e}_x)/2$, $\vec{e}_v = (\vec{e}_t + \vec{e}_x)/2$, and draw $\vec{e}_u$ and $\vec{e}_v$ in a spacetime diagram of the $t - x$ plane.

(b) Show that $\{\vec{e}_u, \vec{e}_v, \vec{e}_y, \vec{e}_z\}$ are a basis for vectors in Minkowski space.

(c) Find the components of the metric tensor on this basis.

(d) Show that $\vec{e}_u$ and $\vec{e}_v$ are null and not orthogonal. (They are called a *null basis* for the $t - x$ plane.) (e) Compute the four one-forms $\tilde{d}u$, $\tilde{d}v$, $\mathbf{g}(\vec{e}_u,)$, $\mathbf{g}(\vec{e}_v,)$ in terms of $\tilde{d}t$ and $\tilde{d}x$.

4
Perfect fluids in special relativity

4.1 Fluids

In many interesting situations in astrophysical GR, the source of the gravitational field can be taken to be a perfect fluid as a first approximation. In general, a 'fluid' is a special kind of *continuum*. A continuum is a collection of particles so numerous that the dynamics of individual particles cannot be followed, leaving only a description of the collection in terms of 'average' or 'bulk' quantities: number of particles per unit volume, density of energy, density of momentum, pressure, temperature, *etc.* The behavior of a lake of water, and the gravitational field it generates, does not depend upon where any one particular water molecule happens to be: it depends only on the average properties of huge collections of molecules.

Nevertheless, these properties can vary from point to point in the lake: the pressure is larger at the bottom than at the top, and the temperature may vary as well. The atmosphere, another fluid, has a density that varies with position. This raises the question of how large a collection of particles to average over: it must clearly be large enough so that the individual particles don't matter, but it must be small enough so that it is relatively homogeneous: the average velocity, kinetic energy, and interparticle spacing must be the same everywhere in the collection. Such a collection is called an '*element*'. This is a somewhat imprecise but useful term for a large collection of particles that may be regarded as having a single value for such quantities as density, average velocity, and temperature. If such a collection doesn't exist (e.g. a *very* rarified gas), then the continuum approximation breaks down.

The continuum approximation assigns to each element a value of density, temperature, *etc.* Since the elements are regarded as 'small', this approximation is expressed mathematically by assigning to each *point* a value of density, temperature, *etc*. So a continuum is defined by various fields, having values at each point and at each time.

So far, this notion of a continuum embraces rocks as well as gases. A *fluid* is a continuum that 'flows': this definition is not very precise, and so the division between solids and fluids is not very well defined. Most solids will flow under high enough pressure. What makes a substance rigid? After some thought we should be able to see that rigidity comes from forces *parallel* to the interface between two elements. Two adjacent elements can push and pull on each other, but the continuum won't be rigid unless they can also prevent each other from sliding along their common boundary. A *fluid* is characterized by the weakness of such antislipping forces compared to the direct push–pull force, which is called pressure. A *perfect* fluid is defined as one in which *all* antislipping forces are zero, and the only force between neighboring fluid elements is pressure. We will soon see how to make this mathematically precise.

4.2 Dust : the number– flux vector $\vec{N}$

We will introduce the relativistic description of a fluid with the simplest one: 'dust' is defined to be a collection of particles, all of which are at rest in some one Lorentz frame. It isn't very clear how this usage of the term 'dust' evolved from the other meaning as that substance which is at rest on the windowsill, but it has become a standard usage in relativity.

The number density n

The simplest question we can ask about these particles is: How many are there per unit volume? In their rest frame, this is merely an exercise in counting the particles and dividing by the volume they occupy. By doing this in many small regions we could come up with different numbers at different points, since the particles may be distributed more densely in one area than in another. We define this *number density* to be n:

$$n := \text{number density in the MCRF of the element.} \tag{4.1}$$

What is the number density in a frame $\bar{\mathcal{O}}$ in which the particles are not at rest? They will all have the same velocity v in $\bar{\mathcal{O}}$. If we look at the same particles as we counted up in the rest frame, then there are clearly the same *number* of particles, but they do not occupy the same volume. Suppose they were originally in a rectangular solid of dimension $\Delta x\ \Delta y\ \Delta z$. The Lorentz contraction will reduce this to $\Delta x\, \Delta y\, \Delta z \sqrt{(1-v^2)}$, since lengths in the direction of motion contract but lengths perpendicular do not (Fig. 4.1). Because of this, the number of particles per unit volume is $[\sqrt{(1-v^2)}]^{-1}$ times what it was in the rest frame:

$$\frac{n}{\sqrt{(1-v^2)}} = \left\{\begin{matrix}\text{number density in frame in} \\ \text{which particles have velocity } v\end{matrix}\right\}. \tag{4.2}$$

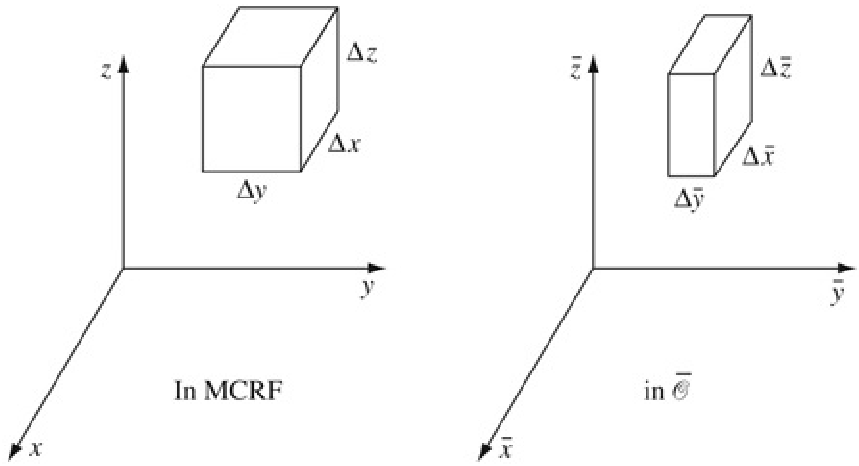


Figure 4.1 The Lorentz contraction causes the density of particles to depend upon the frame in which it is measured.

The flux across a surface

When particles move, another question of interest is, 'how many' of them are moving in a certain direction? This is made precise by the definition of flux: *the flux of particles across a surface is the number crossing a unit area of that surface in a unit time.* This clearly depends on the inertial reference frame ('area' and 'time' are frame-dependent concepts) and on the orientation of the surface (a surface parallel to the velocity of the particles won't be crossed by any of them). In the rest frame of the dust the flux is zero, since all particles are at rest. In the frame $\bar{\mathcal{O}}$, suppose the particles all move with velocity v in the $\bar{x}$ direction, and let us for simplicity consider a surface $\mathcal{S}$ perpendicular to $\bar{x}$ (Fig. 4.2). The rectangular volume outlined by a dashed line clearly contains all and only those particles that will cross the area ΔA of $\mathcal{S}$ in the time $\Delta\bar{t}$. It has volume $v\Delta\bar{t}\ \Delta A$, and contains $[n/\sqrt{(1-v^2)}]v\Delta\bar{t}\ \Delta A$ particles, since in this frame the number density is $n/\sqrt{(1-v^2)}$. The number crossing *per unit* time and per unit area is the flux across surfaces of constant $\bar{x}$: $$(\text{flux})^{\bar{x}} = \frac{nv}{\sqrt{(1-v^2)}}.$$

Suppose, more generally, that the particles had a y component of velocity in $\bar{\mathcal{O}}$ as well. Then the dashed line in Fig. 4.3 encloses all and only those particles that cross ΔA in $\mathcal{S}$ in the time $\Delta\bar{t}$. This is a 'parallelepiped', whose volume is the area of its base times its height. But its height – its extent in the x direction – is just $v^{\bar{x}}\Delta\bar{t}$. Therefore we get

$$(\text{flux})^{\bar{x}} = \frac{nv^{\bar{x}}}{\sqrt{(1-v^2)}}. \tag{4.3}$$

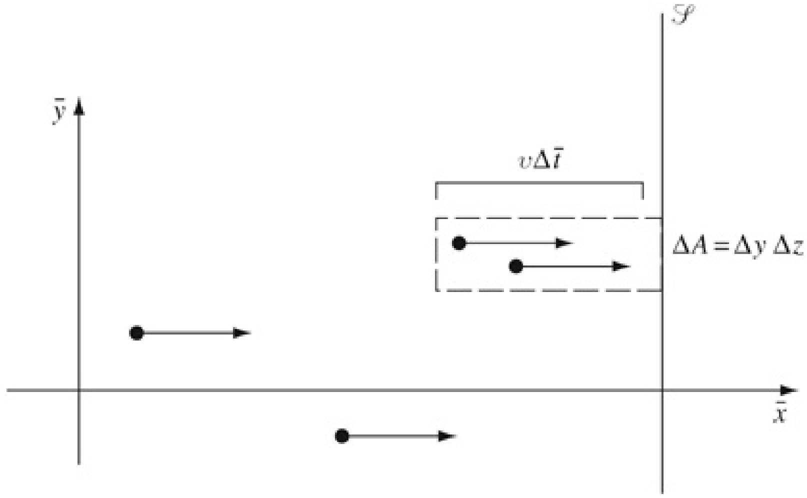


Figure 4.2 Simple illustration of the transformation of flux: if particles move only in the *x*-direction, then all those within a distance $v\Delta\bar{t}$ of the surface $\mathcal{S}$ will cross $\mathcal{S}$ in the time $\Delta\bar{t}$

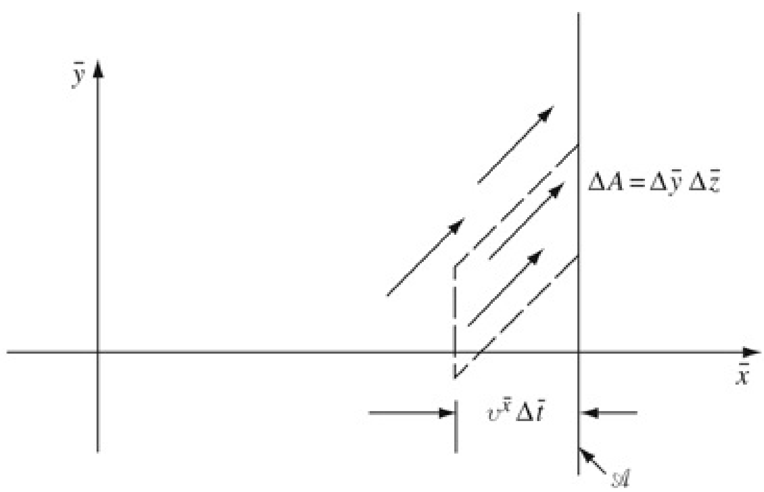


Figure 4.3 The general situation for flux: only the *x*-component of the velocity carries particles across a surface of constant *x*.

The number–flux four-vector $\vec{N}$

Consider the vector $\vec{N}$ defined by

$$\vec{N} = n\vec{U}, \tag{4.4}$$

where $\vec{U}$ is the four-velocity of the particles. In a frame $\bar{\mathcal{O}}$ in which the particles have a velocity (v^x, v^y, v^z), we have $\vec{U} \underset{\bar{\mathcal{O}}}{\to} \left(\frac{1}{\surd(1-v^2)}, \frac{v^x}{\surd(1-v^2)}, \frac{v^y}{\surd(1-v^2)}, \frac{v^z}{\surd(1-v^2)} \right).$

It follows that

$$\vec{N} \underset{\bar{\mathcal{O}}}{\to} \left(\frac{n}{\surd(1-v^2)}, \frac{nv^x}{\surd(1-v^2)}, \frac{nv^y}{\surd(1-v^2)}, \frac{nv^z}{\surd(1-v^2)} \right). \tag{4.5}$$

Thus, in any frame, the time component of $\vec{N}$ is the number density and the spatial components are the fluxes across surfaces of the various coordinates. This is a very important conceptual result. In Galilean physics, number density was a scalar, the same in all frames (no Lorentz contraction), while flux was quite another thing: a three-vector that was frame *dependent*, since the velocities of particles are a frame-dependent notion. Our relativistic approach has unified these two notions into a single, frame-independent four-vector. This is progress in our thinking, of the most fundamental sort: the union of apparently disparate notions into a single coherent one.

It is worth reemphasizing the sense in which we use the word 'frame-independent'. The vector $\vec{N}$ is a geometrical object whose existence is independent of any frame; as a tensor, its action on a one-form to give a number is independent of any frame. Its components *do* of course depend on the frame. Since prerelativity physicists regarded the flux as a three-vector, they had to settle for it as a frame-dependent vector, in the following sense. As a three-vector it was independent of the orientation of the spatial axes in the same sense that four-vectors are independent of all frames; but the flux three-vector is different in frames that move relative to one another, since the velocity of the particles is different in different frames. To the old physicists, a flux vector had to be defined relative to some inertial frame. To a relativist, there is only *one* four-vector, and the frame dependence of the older way of looking at things came from concentrating only on a set of three of the four components of $\vec{N}$. This unification of the Galilean frame-independent number density and frame-dependent flux into a single frame-independent four-vector $\vec{N}$ is similar to the unification of 'energy' and 'momentum' into four-momentum.

One final note: it is clear that

$$\vec{N} \cdot \vec{N} = -n^2, \quad n = (-\vec{N} \cdot \vec{N})^{1/2}. \tag{4.6}$$

Thus, n is a scalar. In the same way that ‘rest mass’ is a scalar, even though energy and ‘inertial mass’ are frame dependent, here we have that n is a scalar, the ‘rest density’, even though number density is frame dependent. We will *always* define n to be a scalar number equal to the number density in the MCRF. We will make similar definitions for pressure, temperature, and other quantities characteristic of the fluid. These will be discussed later.

4.3 One-forms and surfaces

Number density as a timelike flux

We can complete the above discussion of the unity of number density and flux by realizing that number density can be regarded as a timelike flux. To see this, let us look at the flux across x surfaces again, this time in a *spacetime* diagram, in which we plot only $\bar{t}$ and $\bar{x}$ (Fig. 4.4). The surface $\mathcal{S}$ perpendicular to $\bar{x}$ has the world line shown. At any time $\bar{t}$ it is just one point, since we are suppressing both $\bar{y}$ and $\bar{z}$. The world lines of those particles that go through $\mathcal{S}$ in the time $\Delta\bar{t}$ are also shown. The flux is the number of world lines that cross $\mathcal{S}$ in the interval $\Delta\bar{t} = 1$. Really, since it is a two-dimensional surface, its ‘world path’ is three-dimensional, of which we have drawn only a section. The *flux* is the number of world lines that cross a unit ‘volume’ of this three-surface: by volume we of course mean a cube of unit side, $\Delta\bar{t} = 1$, $\Delta\bar{y} = 1$, $\Delta\bar{z} = 1$. So we can define a flux as the number of world lines crossing a unit three-volume. There is no reason we cannot now define this three-volume to be an ordinary spatial volume $\Delta\bar{x} = 1$, $\Delta\bar{y} = 1$, $\Delta\bar{z} = 1$, taken at some particular time $\bar{t}$. This is shown in Fig. 4.5. Now the flux is the number crossing in the interval $\Delta\bar{x} = 1$ (since $\bar{y}$ and $\bar{z}$ are suppressed). But this is just the number ‘contained’ in the unit volume at the given time: the number density. So the ‘timelike’ flux is the number density.

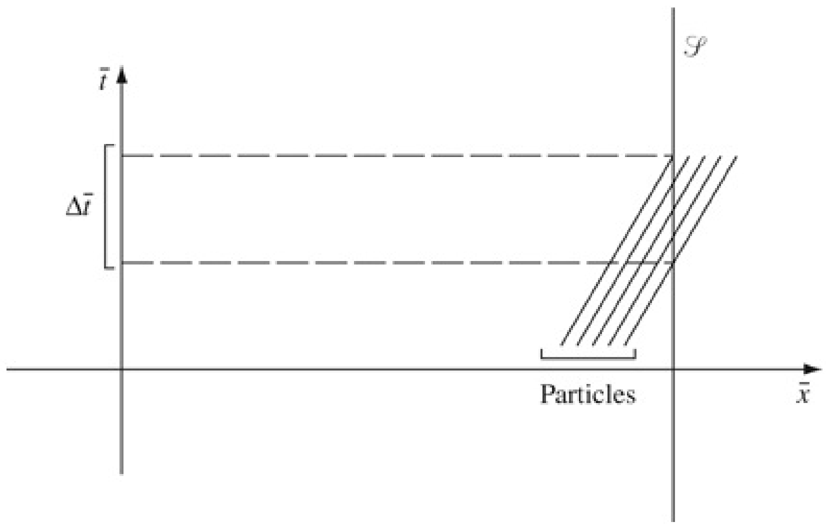


Figure 4.4 Fig. 4.2 in a spacetime diagram, with the $\bar{y}$ direction suppressed.

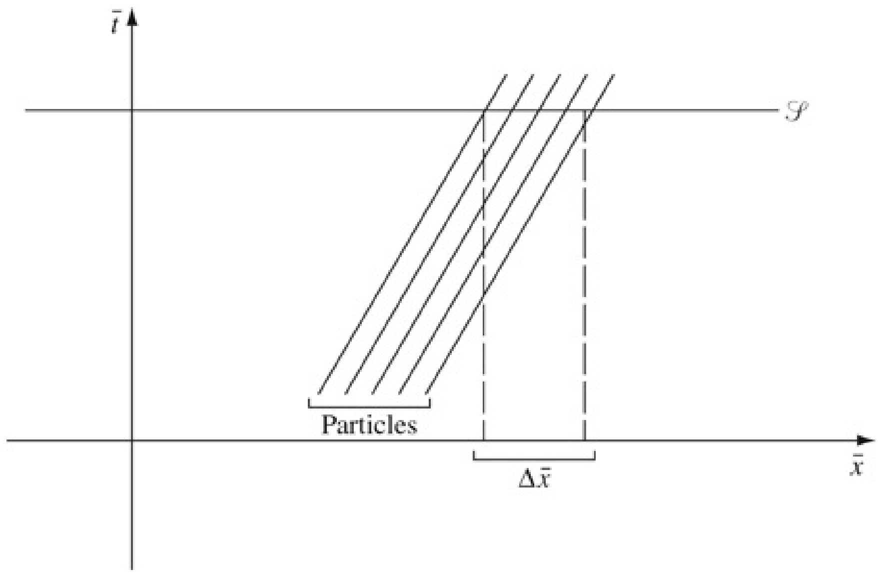


Figure 4.5 Number density as a flux across surfaces $\bar{t}$ = const.

A one-form defines a surface

The way we described surfaces above was somewhat clumsy. To push our invariant picture further we need a somewhat more satisfactory mathematical representation of the surface that these world lines are crossing. This representation is given by one-forms. In general, a surface is defined as the solution to some equation

$$\phi(t, x, y, z) = \text{const.}$$

The gradient of the function ϕ, $\tilde{d}\phi$, is a normal one-form. In some sense, $\tilde{d}\phi$

defines the surface ϕ = const., since it uniquely determines the directions normal to that surface. However, any multiple of $\tilde{d}\phi$ also defines the same surface, so it is customary to use the unit-normal one-form when the surface is not null:

$$\tilde{n} := \tilde{d}\phi/|\tilde{d}\phi|, \tag{4.7}$$

where

$$\begin{aligned}&|\tilde{d}\phi| \text{ is the magnitude of } \tilde{d}\phi : \\ &|\tilde{d}\phi| = |\,\eta^{\alpha\beta}\,\phi_{,\alpha}\,\phi_{,\beta}\,|^{1/2}.\end{aligned} \tag{4.8}$$

(Do not confuse $\tilde{n}$ with n, the number density in the MCRF: they are completely different, given, by historical accident, the same letter.)

As in three-dimensional vector calculus (e.g. Gauss' law), we define the 'surface element' as the unit normal times an area element in the surface. In this case, a volume element in a three-space whose coordinates are x^α, x^β, and x^γ (for some *particular* values of α, β, and γ, all distinct) can be represented by

$$\tilde{n}\,dx^\alpha\,dx^\beta\,dx^\gamma, \tag{4.9}$$

and a *unit* volume ($dx^\alpha = dx^\beta = dx^\gamma = 1$) is just $\tilde{n}$. (These dxs *are* the infinitesimals that we integrate over, not the gradients.) The flux across the surface

Recall from Gauss' law in three dimensions that the flux across a surface of, say, the electric field is just $\mathbf{E}\cdot\mathbf{n}$, the dot product of $\mathbf{E}$ with the unit normal. The situation here is exactly the same: the flux (of particles) across a surface of constant ϕ is $\langle\tilde{n}, \vec{N}\rangle$. To see this, let ϕ be a coordinate, say $\bar{x}$. Then a surface of constant $\bar{x}$ has normal $\tilde{d}\bar{x}$, which is a unit normal already since $\tilde{d}\bar{x} \rightarrow_{\bar{\mathcal{O}}} (0, 1, 0, 0)$. Then $\langle\tilde{d}\bar{x}, \vec{N}\rangle = N^\alpha(\tilde{d}\bar{x})_\alpha = N^{\bar{x}}$, which is what we have already seen is the flux across $\bar{x}$ surfaces. Clearly, had we chosen $\phi = \bar{t}$, then we would have wound up with $N^{\bar{0}}$, the number density, or flux across a surface of constant $\bar{t}$.

This is one of the first concrete physical examples of our definition of a vector as a function of one-forms into real numbers. Given the vector $\vec{N}$, we can calculate the flux across a surface by finding the unit-normal one-form for that surface, and contracting it with $\vec{N}$. We have, moreover, expressed everything frame invariantly and in a manner that separates the property of the system of particles $\vec{N}$ from the property of the surface $\tilde{n}$. All of this will have many parallels in § 4.4 below.

Representation of a frame by a one-form

Before going on to discuss other properties of fluids, we should mention a useful fact. An inertial frame, which up to now has been defined by its four-velocity, can be defined also by a one-form, namely that associated with its four-velocity $\mathbf{g}(\vec{U},\ )$. This has components

$$U_\alpha = \eta_{\alpha\beta}\, U^\beta$$

or, in this frame,

$$U_0 = -1, U_i = 0.$$

This is clearly also equal to $-\tilde{\mathrm{d}}t$ (since their components are equal). So we could equally well define a frame by giving $\tilde{\mathrm{d}}t$. This has a nice picture: $\tilde{\mathrm{d}}t$ is to be pictured as a set of surfaces of constant t, the surfaces of simultaneity. These clearly *do* define the frame, up to spatial rotations, which we usually ignore. In fact, in some sense $\tilde{\mathrm{d}}t$ is a more natural way to define the frame than $\vec{U}$. For instance, the energy of a particle whose four-momentum is $\vec{p}$ is

$$E = \langle \tilde{\mathrm{d}}t, \vec{p}\,\rangle = p^0. \tag{4.10}$$

There is none of the awkward minus sign that we get in Eq. (2.35)

$$E = -\vec{p}\cdot\vec{U}.$$

4.4 Dust again : the stress–energy tensor

So far we have only discussed how many dust particles there are. But they also have energy and momentum, and it will turn out that their energy and momentum are the source of the gravitational field in GR. So we must now ask how to represent them in a frame-invariant manner. We will assume for simplicity that all the dust particles have the same rest mass m.

Energy density

In the MCRF, the energy of each particle is just m, and the number per unit volume is n. Therefore the energy per unit volume is mn. We denote this in general by ρ:

$$\rho := \text{energy density in the MCRF.} \tag{4.11}$$

Thus ρ is a scalar just as n is (and m is). In our case of dust,

$$\rho = nm \text{ (dust)}. \tag{4.12}$$

In more general fluids, where there is random motion of particles and hence kinetic energy of motion, even in an average rest frame, Eq. (4.12) will not be valid.

In the frame $\bar{\mathcal{O}}$ we again have that the number density is $n/\sqrt{(1 - v^2)}$, but now the energy of each particle is $m/\sqrt{(1 - v^2)}$, since it is moving. Therefore the energy density is $mn/(1 - v^2)$:

$$\frac{\rho}{1 - v^2} = \left\{ \begin{matrix} \text{energy density in a frame in} \\ \text{which particles have velocity } \mathbf{v} \end{matrix} \right\}. \tag{4.13}$$

This transformation involves *two* factors of $(1 - v^2)^{-1/2} = \Lambda^{\bar{0}}{}_0$, because *both* volume *and* energy transform. It is impossible, therefore, to represent energy density as some component of a vector. It is, in fact, a component of a $\binom{2}{0}$ tensor. This is most easily seen from the point of view of our definition of a tensor. To define energy requires a one-form, in order to select the zero component of the four-vector of energy and momentum; to define a density also requires a one-form, since density is a flux across a constant-time surface. Similarly, an energy flux also requires two one-forms: one to define 'energy' and the other to define the surface. We can also speak of momentum density: again a one-form defines which component of momentum, and another one-form defines density. By analogy there is also momentum flux: the rate at which momentum crosses some surface. All these things require two one-forms. So there is a tensor **T**, called the stress–energy tensor, which has all these numbers as values when supplied with the appropriate one-forms as arguments.

Stress–energy tensor

The most convenient definition of the stress–energy tensor is in terms of its components in some (arbitrary) frame:

$$\mathbf{T}(\tilde{\mathrm{d}}x^\alpha, \tilde{\mathrm{d}}x^\beta) = T^{\alpha\beta} := \left\{ \begin{matrix} \text{flux of } \alpha \text{ momentum across} \\ \text{a surface of constant } x^\beta \end{matrix} \right\}. \tag{4.14}$$

(By α momentum we mean, of course, the α component of four-momentum: $p^\alpha := \langle \tilde{\mathrm{d}}x^\alpha, \vec{p} \rangle$.) That this is truly a tensor is proved in Exer. 5, § 4.10.

Let us see how this definition fits in with our discussion above. Consider T^{00}. This is defined as the flux of zero momentum (energy) across a surface t = constant. This is just the energy density:

$$T^{00} = \text{energy density}. \tag{4.15}$$

Similarly, T^{0i} is the flux of energy across a surface x^i = const:

$$T^{0i} = \text{energy flux across } x^i \text{ surface}. \tag{4.16}$$

Then T^{i0} is the flux of i momentum across a surface t = const: the density of i momentum,

$$T^{i0} = i \text{ momentum density}. \tag{4.17}$$

Finally, T^{ij} is the j flux of i momentum:

$$T^{ij} = \text{flux of } i \text{ momentum across } j \text{ surface}. \tag{4.18}$$

For any particular system, giving the components of **T** in some frame defines it completely. For dust, the components of **T** in the MCRF are particularly easy. There is no motion of the particles, so all i momenta are zero and all spatial fluxes are zero. Therefore

$$(T^{00})_{\mathrm{MCRF}} = \rho = mn,$$
$$(T^{0i})_{\mathrm{MCRF}} = (T^{i0})_{\mathrm{MCRF}} = (T^{ij})_{\mathrm{MCRF}} = 0.$$

It is easy to see that the tensor $\vec{p} \otimes \vec{N}$ has exactly these components in the MCRF, where $\vec{p} = m\vec{U}$ is the four-momentum of a particle. Therefore we have

$$\text{Dust}: \mathbf{T} = \vec{p} \otimes \vec{N} = mn\, \vec{U} \otimes \vec{U} = \rho\, \vec{U} \otimes \vec{U}. \tag{4.19}$$

From this we can conclude

$$\begin{aligned} T^{\alpha\beta} &= \mathbf{T}(\tilde{\omega}^\alpha, \tilde{\omega}^\beta) \\ &= \rho \vec{U}(\tilde{\omega}^\alpha) \vec{U}(\tilde{\omega}^\beta) \\ &= \rho U^\alpha U^\beta. \end{aligned} \tag{4.20}$$

In the frame $\bar{\mathcal{O}}$, where

$$\vec{U} \to \left(\frac{1}{\sqrt{(1 - v^2)}}, \frac{v^x}{\sqrt{(1 - v^2)}}, \dots \right),$$

we therefore have

$$\left.\begin{aligned} T^{\bar{0}\bar{0}} &= \rho U^{\bar{0}} U^{\bar{0}} = \rho/(1 - v^2), \\ T^{\bar{0}\bar{i}} &= \rho U^{\bar{0}} U^{\bar{i}} = \rho v^i/(1 - v^2), \\ T^{\bar{i}\bar{0}} &= \rho U^{\bar{i}} U^{\bar{0}} = \rho v^i(1 - v^2), \\ T^{\bar{i}\bar{j}} &= \rho U^{\bar{i}} U^{\bar{j}} = \rho v^i v^j/(1 - v^2). \end{aligned}\right\} \tag{4.21}$$

These are exactly what we would calculate, from first principles, for energy density, energy flux, momentum density, and momentum flux respectively. (We did the calculation for energy density above.) Notice one important point: $T^{\alpha\beta} = T^{\beta\alpha}$; that is, **T** is symmetric. This will turn out to be true in general, not just for dust.

4.5 General fluids

Until now we have dealt with the simplest possible collection of particles. To generalize this to real fluids, we have to take account of the facts that (i) besides the bulk motions of the fluid, each particle has some random velocity; and (ii) there may be various forces between particles that contribute potential energies to the total.

Definition of macroscopic quantities

The concept of a fluid element was discussed in § 4.1. For each fluid element, we go to the frame in which it is at rest (its total spatial momentum is zero). This is its MCRF. This frame is truly *momentarily* comoving: since fluid elements can be accelerated, a moment later a different inertial frame will be the MCRF. Moreover, two different fluid elements may be moving relative to one another, so that they would not have the same MCRFs. Thus, the MCRF is specific to a single fluid element, and which frame is the MCRF is a function of position and time. *All scalar quantities associated with a fluid element in relativity* (such as number density, energy density, and temperature) *are defined to be their values in the MCRF*. Thus we make the definitions displayed in Table 4.1. We confine our attention to fluids that consist of only one component, one kind of particle, so that (for example) interpenetrating flows are not possible.

Table 4.1 Macroscopic quantities for single-component fluids

Symbol	Name	Definition
$\vec{U}$	Four-velocity of fluid element	Four-velocity of MCRF
n	Number density	Number of particles per unit volume in MCRF
$\vec{N}$	Flux vector	$\vec{N} := n\vec{U}$
ρ	energy density	Density of *total* mass energy (rest mass, random kinetic, chemical, ...)
Π	Internal energy per particle	$\Pi := (\rho/n) - m \Rightarrow \rho = n(m + \Pi)$ Thus Π is a general name for all energies other than the rest mass.
ρ_0	Rest-mass density	$\rho_0 := mn$. Since m is a constant, this is the 'energy' associated with the rest mass only. Thus, $\rho = \rho_0 + n\Pi$.
T	Temperature	Usual thermodynamic definition in MCRF (see below).
p	Pressure	Usual fluid-dynamical notion in MCRF. More about this later.
S	Specific entropy	Entropy per particle (see below).

First law of thermodynamics

This law is simply a statement of conservation of energy. In the MCRF, we imagine that the fluid element is able to exchange energy with its surroundings in only two ways: by heat conduction (absorbing an amount of heat ΔQ) and by work (doing an amount of work $p\Delta V$, where V is the three-volume of the element). If we let E be the total energy of the element, then since ΔQ is energy gained and $p\Delta V$ is energy lost, we can write (assuming small changes)

$$\left.\begin{array}{l} \Delta E = \Delta Q - p\Delta V, \\ \text{or} \\ \Delta Q = \Delta E + p\Delta V. \end{array}\right\} \tag{4.22}$$

Now, if the element contains a total of N particles, and if this number doesn't change (i.e. no creation or destruction of particles), we can write

(4.23)

$$V = \frac{N}{n}, \quad \Delta V = -\frac{N}{n^2}\Delta n. \tag{4.23}$$

Moreover, we also have (from the definition of ρ)

$$E = \rho V = \rho N/n,$$
$$\Delta E = \rho \Delta V + V \Delta \rho.$$

These two results imply

$$\Delta Q = \frac{N}{n}\Delta \rho - N(\rho + p)\frac{\Delta n}{n^2}.$$

If we write $q := Q/N$, which is the heat absorbed per particle, we obtain

$$n\,\Delta q = \Delta \rho - \frac{\rho + p}{n}\Delta n. \tag{4.24}$$

Now suppose that the changes are 'infinitesimal'. It can be shown in general that a fluid's state can be given by two parameters: for instance, ρ and T or ρ and n. Everything else is a function of, say, ρ and n. That means that the right-hand side of Eq. (4.24), $\mathrm{d}\rho - (\rho + p)\mathrm{d}n/n,$

depends only on ρ and n. The general theory of first-order differential equations shows that this *always* possesses an *integrating factor*: that is, there exist two functions A and B, functions only of ρ and n, such that $\mathrm{d}\rho - (\rho + p)\mathrm{d}n/n \equiv A\,\mathrm{d}B$

is an identity for all ρ and n. It is customary in thermodynamics to *define* temperature to be A/n and specific entropy to be B:

$$\mathrm{d}\rho - (\rho + p)\,\mathrm{d}n/n = nT\,\mathrm{d}S, \tag{4.25}$$

or, in other words,

$$\Delta q = T\,\Delta S. \tag{4.26}$$

The heat absorbed by a fluid element is proportional to its increase in entropy.

We have thus introduced T and S as convenient mathematical definitions. A full treatment would show that T is the thing normally meant by temperature, and that S is the thing used in the second law of thermodynamics, which says that the *total* entropy in any system must increase. We'll have nothing to say about the second law. Entropy appears here only because it is an integral of the first law, which is merely conservation of energy. In particular, we shall use both Eqs. (4.25) and (4.26) later.

The general stress–energy tensor

The definition of $T^{\alpha\beta}$ in Eq. (4.14) is perfectly general. Let us in particular look at it in the MCRF, where there is no bulk flow of the fluid element, and no spatial momentum in the particles. Then *in the MCRF* we have:

) T^{00} = energy density = ρ.

) T^{0i} = energy flux. Although there is no motion in the MCRF, energy may be transmitted by heat conduction. So T^{0i} is basically a heat-conduction term in the MCRF.

) T^{i0} = momentum density. Again the particles themselves have no net momentum in the MCRF, but if heat is being conducted, then the moving energy will have an associated momentum. We'll argue below that $T^{i0} \equiv T^{0i}$.

) T^{ij} = momentum flux. This is an interesting and important term. The next section gives a thorough discussion of it. It is called the *stress*.

The spatial components of *T*, T^{ij}

By definition, T^{ij} is the flux of i momentum across the j surface. Consider (Fig. 4.6) two adjacent fluid elements, represented as cubes, having the common interface $\mathcal{S}$. In general, they exert forces on each other. Shown in the diagram is the force **F** exerted by A on B (B of course exerts an equal and opposite force on A). Since force equals the rate of change of momentum (by Newton's law, which is valid here, since we are in the MCRF where velocities are zero), A is pouring momentum into B at the rate **F** per unit time. Of course, B may or may not acquire a new velocity as a result of this new momentum it acquires; this depends upon how much momentum is put into B by its other neighbors. Obviously B's motion is the resultant of all the forces. Nevertheless, each force adds momentum to B. There is therefore a flow of momentum across $\mathcal{S}$ from A to B at the rate **F**. If $\mathcal{S}$ has area $\mathcal{A}$, then the flux of momentum across $\mathcal{S}$ is $\mathbf{F}/\mathcal{A}$. If $\mathcal{S}$ is a surface of constant x^j, then T^{ij} for fluid element A is $F^i/\mathcal{A}$.

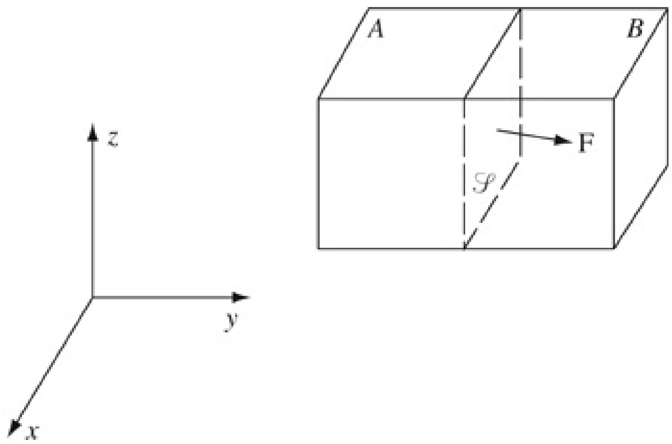


Figure 4.6 The force **F** exerted by element *A* on its neighbor *B* may be in any direction depending on properties of the medium and any external forces.

This is a brief illustration of the meaning of T^{ij}: it represents forces between adjacent fluid elements. As mentioned before, these forces need not be perpendicular to the surfaces between the elements (i.e. viscosity or other kinds of rigidity give forces parallel to the interface). But if the forces *are* perpendicular to the interfaces, then T^{ij} will be zero unless $i = j$. (Think this through – we'll use it shortly.) Symmetry of $T^{\alpha\beta}$ in MCRF

We now prove that **T** is a symmetric tensor. We need only prove that its components are symmetric in one frame; that implies that for any $\tilde{r}$, $\tilde{q}$, $\mathbf{T}(\tilde{r}, \tilde{q}) = \mathbf{T}(\tilde{q}, \tilde{r})$, which implies the symmetry of its components in any other frame. The easiest frame is the MCRF.

(a) Symmetry of T^{ij}. Consider Fig. 4.7 in which we have drawn a fluid element as a cube of side *l*. The force it exerts on a neighbor across surface (1) (a surface x = const.) is $F_1^i = T^{ix}l^2$, where the factor l^2 gives the area of the face. Here, i runs over 1, 2, and 3, since **F** is not necessarily perpendicular to the surface. Similarly, the force it exerts on a neighbor across (2) is $F_2^i = T^{iy}l^2$. (We shall take the limit $l \to 0$, so bear in mind that the element is small.) The element also exerts a force on its neighbor toward the $-x$ direction, which we call F_3^i. Similarly, there is F_4^i on the face looking in the negative y direction. The forces *on* the fluid element are, respectively, $-F_1^i, -F_2^i$, *etc.* The first point is that $F_3^i \approx -F_1^i$ in order that the sum of the forces on the element should vanish when $l \to 0$ (otherwise the tiny mass obtained as $l \to 0$ would have an infinite acceleration). The next point is to compute torques about the z axis through the center of the fluid element. (Since forces on the top and bottom of the cube don't contribute to this, we haven't considered them.) For the torque calculation it is convenient to place the origin of coordinates at the center of the cube. The torque due to $-(\mathbf{r} \times \mathbf{F}_1)^z = -xF_1^y = -\frac{1}{2}lT^{yx}l^2$, where we have approximated the force as acting at the center of the face, where $\mathbf{r} \to (l/2, 0, 0)$ (note particularly that $y = 0$ there). The torque due to $-\mathbf{F}_3$ is the *same*, $-\frac{1}{2}l^3T^{yx}$. The torque due to $-\mathbf{F}_2$ is $-(\mathbf{r} \times \mathbf{F}_2)^z = +yF_2^x = \frac{1}{2}lT^{xy}l^2$. Similarly, the torque due to $-\mathbf{F}_4$ is the same, $\frac{1}{2}l^3T^{xy}$. Therefore, the total torque is

$$\tau_z = l^3(T^{xy} - T^{yx}). \tag{4.27}$$

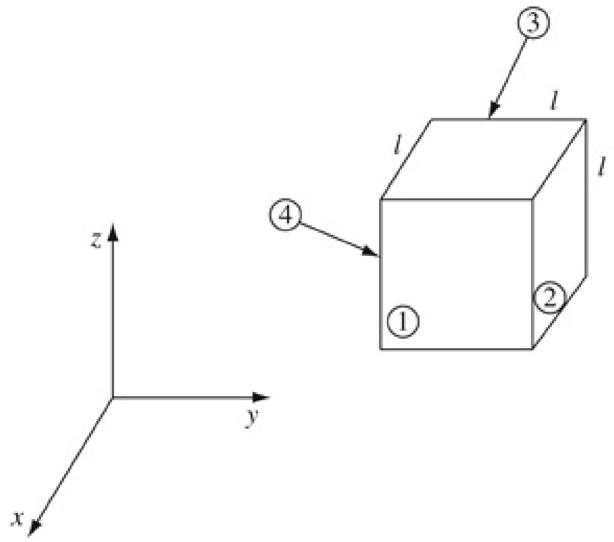


Figure 4.7 A fluid element.

The moment of inertia of the element about the z axis is proportional to its mass times l^2, or

$$I = \alpha \rho l^5,$$

where α is some numerical constant and ρ is the density (whether of total energy or rest mass doesn't matter in this argument). Therefore the angular acceleration is

$$\ddot{\theta} = \frac{\tau}{I} = \frac{T^{xy} - T^{yx}}{\alpha \rho l^2}. \tag{4.28}$$

Since α is a number and ρ is independent of the size of the element, as are T^{xy} and T^{yx}, this will go to infinity as $l \to 0$ unless $T^{xy} = T^{yx}$.

Thus, since it is obviously not true that fluid elements are whirling around inside fluids, smaller ones whirling ever faster, we have that the stresses are always *symmetric*:

$$T^{ij} = T^{ji}. \tag{4.29}$$

Since we made no use of any property of the substance, this is true of solids as well as fluids. It is true in Newtonian theory as well as in relativity; in Newtonian theory T^{ij} are the components of a three-dimensional $\binom{2}{0}$ tensor called the stress tensor. It is familiar to 0 any materials engineer; and it contributes its name to its relativistic generalization **T**.

(b) Equality of momentum density and energy flux. This is much easier to demonstrate. The energy flux is the density of energy times the speed it flows at. But since energy and mass are the same, this is the density of mass times the speed it is moving at; in other words, the density of momentum. Therefore $T^{0i} = T^{i0}$.

Conservation of energy–momentum

Since **T** represents the energy and momentum content of the fluid, there must be some way of using it to express the law of conservation of energy and momentum. In fact it is reasonably easy. In Fig. 4.8 we see a cubical fluid element, seen only in cross-section (z direction suppressed). Energy can flow in across all sides. The rate of flow across face (4) is $l^2T^{0x}(x = 0)$, and across (2) is $-l^2T^{0x}(x = a)$; the second term has a minus sign, since T^{0x} represents energy flowing in the positive x direction, which is out of the volume across face (2). Similarly, energy flowing in the y direction is $l^2T^{0y}(y = 0) - l^2T^{0y}(y = l)$. The sum of these rates must be the rate of increase in the energy inside, $\partial(T^{00}l^3)/\partial t$ (statement of conservation of energy). Therefore we have

$$\begin{aligned}\frac{\partial}{\partial t}l^3T^{00} = l^2\Big[&T^{0x}(x=0) - T^{0x}(x=l) + T^{0y}(y=0)\\ &-T^{oy}(y=l) + T^{0z}(z=0) - T^{0z}(z=l)\Big].\end{aligned} \tag{4.30}$$

Dividing by l^3 and taking the limit $l \to 0$ gives

$$\frac{\partial}{\partial t}T^{00} = -\frac{\partial}{\partial x}T^{0x} - \frac{\partial}{\partial y}T^{0y} - \frac{\partial}{\partial z}T^{0z}. \tag{4.31}$$

[In deriving this we use the definition of the derivative

$$\lim_{l\to 0}\frac{T^{0x}(x=0)-T^{0x}(x=l)}{l} \equiv -\frac{\partial}{\partial x}T^{0x}.] \tag{4.32}$$

Eq. (4.31) can be written as

$$T^{00}{}_{,0} + T^{0x}{}_{,x} + T^{0y}{}_{,y} + T^{0z}{}_{,z} = 0$$

or

$$T^{0\alpha}{}_{,\alpha} = 0. \tag{4.33}$$

This is the statement of the law of conservation of energy.

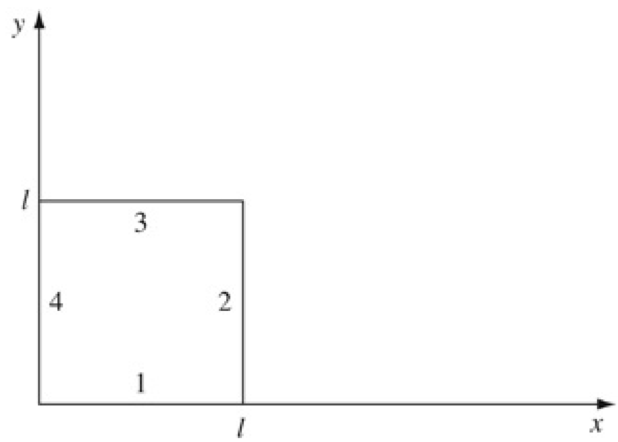


Figure 4.8 A section z = const. of a cubical fluid element.

Similarly, momentum is conserved. The same mathematics applies, with the index ‘0’ changed to whatever spatial index corresponds to the component of momentum whose conservation is being considered. The general conservation law is, then,

$$T^{\alpha\beta}{}_{,\beta} = 0. \tag{4.34}$$

This applies to any material in SR. Notice it is just a four-dimensional divergence. Its relation to Gauss’ theorem, which gives an integral form of the conservation law, will be discussed later.

Conservation of particles

It may also happen that, during any flow of the fluid, the number of particles in a fluid element will change, but of course the total number of particles in the fluid will not change. In particular, in Fig. 4.8 the rate of change of the number of particles in a fluid element will be due only to loss or gain across the boundaries, *i.e.* to net fluxes out or in. This conservation law is derivable in the same way as Eq. (4.34) was. We can then write that

$$\frac{\partial}{\partial t}N^0 = -\frac{\partial}{\partial x}N^x - \frac{\partial}{\partial y}N^y - \frac{\partial}{\partial z}N^z$$

or

$$N^\alpha{}_{,\alpha} = (nU^\alpha)_{,\alpha} = 0. \tag{4.35}$$

We will confine ourselves to discussing only fluids that obey this conservation law. This is hardly any restriction, since n can, if necessary, always be taken to be the density of baryons.

'Baryon', for those not familiar with high-energy physics, is a general name applied to the more massive particles in physics. The two commonest are the neutron and proton. All others are too unstable to be important in everyday physics – but when they decay they form protons and neutrons, thus conserving the total number of baryons without conserving rest mass or particle identity. Although theoretical physics suggests that baryons may not always be conserved – for instance, so-called 'grand unified theories' of the strong, weak, and electromagnetic interactions may predict a finite lifetime for the proton, and collapse to and subsequent evaporation of a black hole (see Ch. 11) will not conserve baryon number – no such phenomena have yet been observed and, in any case, are unlikely to be important in most situations.

4.6 Perfect fluids

Finally, we come to the type of fluid which is our principal subject of interest. *A perfect fluid in relativity is defined as a fluid that has no viscosity and no heat* *conduction in the MCRF*. It is a generalization of the 'ideal gas' of ordinary thermodynamics. It is, next to dust, the simplest kind of fluid to deal with. The two restrictions in its definition simplify enormously the stress–energy tensor, as we now see.

No heat conduction

From the definition of **T**, we see that this immediately implies that, in the MCRF, $T^{0i} = T^{i0} = 0$. Energy can flow only if particles flow. Recall that in our discussion of the first law of thermodynamics we showed that if the number of particles was conserved, then the specific entropy was related to heat flow by Eq. (4.26). This means that in a perfect fluid, if Eq. (4.35) for conservation of particles is obeyed, then we should also have that S is a constant in time during the flow of the fluid. We shall see how this comes out of the conservation laws in a moment.

No viscosity

Viscosity is a force parallel to the interface between particles. Its absence means that the forces should always be perpendicular to the interface, *i.e.* that T^{ij} should be zero unless $i = j$. This means that T^{ij} should be a diagonal matrix. Moreover, it must be diagonal in *all* MCRF frames, since 'no viscosity' is a statement independent of the spatial axes. The only matrix diagonal in all frames is a multiple of the identity: all its diagonal terms are equal. Thus, an x surface will have across it only a force in the x direction, and similarly for y and z; these forces-per-unit-area are all equal, and are called the *pressure*, p. So we have $T^{ij} = p\delta^{ij}$. From six possible quantities (the number of independent elements in the 3×3 symmetric matrix T^{ij}) the zero-viscosity assumption has reduced the number of functions to one, the pressure.

Form of T

In the MCRF, **T** has the components we have just deduced:

$$(T^{\alpha\beta}) = \begin{pmatrix} \rho & 0 & 0 & 0 \\ 0 & p & 0 & 0 \\ 0 & 0 & p & 0 \\ 0 & 0 & 0 & p \end{pmatrix}. \tag{4.36}$$

It is not hard to show that in the MCRF

$$T^{\alpha\beta} = (\rho + p)U^\alpha U^\beta + p\eta^{\alpha\beta}. \tag{4.37}$$

For instance, if $\alpha = \beta = 0$, then $U^0 = 1$, $\eta^{00} = -1$, and $T^{\alpha\beta} = (\rho + p) - p = \rho$, as in Eq. (4.36). By trying all possible α and β you can verify that Eq. (4.37) gives Eq. (4.36). But Eq. (4.37) is a frame-invariant formula in the sense that it uniquely implies

$$\mathbf{T} = (\rho + p)\vec{U} \otimes \vec{U} + p\mathbf{g}^{-1}. \tag{4.38}$$

This is the stress–energy tensor of a perfect fluid.

Aside on the meaning of pressure

A comparison of Eq. (4.38) with Eq. (4.19) shows that ‘dust’ is the special case of a pressure-free perfect fluid. This means that a perfect fluid can be pressure free only if its particles have *no* random motion at all. Pressure arises in the random velocities of the particles. Even a gas so dilute as to be virtually collisionless has pressure. This is because pressure is the flux of momentum; whether this comes from forces or from particles crossing a boundary is immaterial.

The conservation laws

Eq. (4.34) gives us

$$T^{\alpha\beta}{}_{,\beta} = \left[(\rho + p)U^\alpha U^\beta + p\eta^{\alpha\beta}\right]_{,\beta} = 0. \tag{4.39}$$

This gives us our first real practice with tensor calculus. There are four equations in Eq. (4.39), one for each α. First, let us also assume

$$(nU^\beta)_{,\beta} = 0 \tag{4.40}$$

and write the first term in Eq. (4.39) as

$$[(\rho + p)U^\alpha U^\beta]_{,\beta} = \left[\frac{\rho + p}{n} U^\alpha n U^\beta\right]_{,\beta}$$
$$= nU^\beta \left(\frac{\rho + p}{n} U^\alpha\right)_{,\beta}. \tag{4.41}$$

Moreover, $\eta^{\alpha\beta}$ is a constant matrix, so $\eta^{\alpha\beta}{}_{,\gamma} = 0$. This also implies, by the way, that

$$U^\alpha{}_{,\beta} U_\alpha = 0. \tag{4.42}$$

The proof of Eq. (4.42) is

$$U^\alpha U_\alpha = -1 \Rightarrow (U^\alpha U_\alpha)_{,\beta} = 0 \tag{4.43}$$

or

$$(U^\alpha U^\gamma \eta_{\alpha\gamma})_{,\beta} = (U^\alpha U^\gamma)_{,\beta} \eta_{\alpha\gamma} = 2U^\alpha{}_{,\beta} U^\gamma \eta_{\alpha\gamma}. \tag{4.44}$$

The last step follows from the symmetry of $\eta_{\alpha\beta}$, which means that $U^\alpha{}_{,\beta} U^\gamma \eta_{\alpha\gamma} = U^\alpha U^\gamma{}_{,\beta} \eta_{\alpha\gamma}$. Finally, the last expression in Eq. (4.44) converts to $2U^\alpha{}_{,\beta} U_\alpha$,

which is zero by Eq. (4.43). This proves Eq. (4.42). We can make use of Eq. (4.42) in the following way. The original equation now reads, after use of Eq. (4.41),

$$nU^\beta \left(\frac{\rho + p}{n} U^\alpha\right)_{,\beta} + p_{,\beta} \eta^{\alpha\beta} = 0. \tag{4.45}$$

From the four equations here, we can obtain a particularly useful one. Multiply by U_α and sum on α. This gives the time component of Eq. (4.45) in the MCRF:

$$nU^\beta U_\alpha \left(\frac{\rho + p}{n} U^\alpha\right)_{,\beta} + p_{,\beta} \eta^{\alpha\beta} U_\alpha = 0. \tag{4.46}$$

The last term is just

$$p_{,\beta} U^\beta,$$

which we know to be the derivative of p along the world line of the fluid element, $dp/d\tau$. The first term gives zero when the β derivative operates on U^α

(by Eq. 4.42), so we obtain (using $U^\alpha U_\alpha = -1$)

$$U^\beta \left[-n \left(\frac{\rho + p}{n} \right)_{,\beta} + p_{,\beta} \right] = 0. \tag{4.47}$$

A little algebra converts this to

$$-U^\beta \left[\rho_{,\beta} - \frac{\rho + p}{n} n_{,\beta} \right] = 0. \tag{4.48}$$

Written another way,

$$\frac{\mathrm{d}\rho}{\mathrm{d}\tau} - \frac{\rho + p}{n} \frac{\mathrm{d}n}{\mathrm{d}\tau} = 0. \tag{4.49}$$

This is to be compared with Eq. (4.25). It means

$$U^\alpha S_{,\alpha} = \frac{\mathrm{d}S}{\mathrm{d}\tau} = 0. \tag{4.50}$$

Thus, the flow of a particle-conserving perfect fluid conserves specific entropy. This is called *adiabatic*. Because entropy is constant in a fluid element as it flows, we shall not normally need to consider it. Nevertheless, it is important to remember that the law of conservation of energy in thermodynamics is embodied in the component of the conservation equations, Eq. (4.39), parallel to U^α.

The remaining three components of Eq. (4.39) are derivable in the following way. We write, again, Eq. (4.45):

$$nU^\beta \left(\frac{\rho + p}{n} U^\alpha \right)_{,\beta} + p_{,\beta} \eta^{\alpha\beta} = 0$$

and go to the MCRF, where $U^i = 0$ but $U^i{}_{,\beta} \neq 0$. In the MCRF, the zero component of this equation is the same as its contraction with U_α, which we have just examined. So we only need the i components:

$$nU^\beta \left(\frac{\rho + p}{n} U^i \right)_{,\beta} + p_{,\beta} \eta^{i\beta} = 0. \tag{4.51}$$

Since $U^i = 0$, the β derivative of $(\rho + p)/n$ contributes nothing, and we get

$$(\rho + p)U^i{}_{,\beta}U^\beta + p_{,\beta}\eta^{i\beta} = 0. \tag{4.52}$$

Lowering the index i makes this easier to read (and changes nothing). Since $\eta_i^{\ \beta} = \delta_i^{\ \beta}$ we get

$$(\rho + p)U_{i,\beta}U^\beta + p_{,i} = 0. \tag{4.53}$$

Finally, we recall that $U_{i,\beta}U^\beta$ is the definition of the four-acceleration a_i:

$$(\rho + p)a_i + p_{,i} = 0. \tag{4.54}$$

Those familiar with nonrelativistic fluid dynamics will recognize this as the generalization of

$$\rho\mathbf{a} + \nabla p = 0, \tag{4.55}$$

where

$$\mathbf{a} = \dot{\mathbf{v}} + (\mathbf{v} \cdot \nabla)\mathbf{v}. \tag{4.56}$$

The only difference is the use of $(\rho + p)$ instead of ρ. In relativity, $(\rho + p)$ plays the role of 'inertial mass density', in that, from Eq. (4.54), the larger $(\rho + p)$, the harder it is to accelerate the object. Eq. (4.54) is essentially $\mathbf{F} = m\mathbf{a}$, with $-p_{,i}$ being the force per unit volume on a fluid element. Roughly speaking, p is the force a fluid element exerts on its neighbor, so $-p$ is the force on the element. But the neighbor on the opposite side of the element is pushing the other way, so only if there is a change in p across the fluid element will there be a net force causing it to accelerate. That is why $-\nabla p$ gives the force.

4.7 Importance for general relativity

General relativity is a relativistic theory of gravity. We weren't able to plunge into it immediately because we lacked a good enough understanding of tensors, of fluids in SR, and of curved spaces. We have yet to study curvature (that comes next), but at this point we can look ahead and discern the vague outlines of the theory we shall study.

The first comment is on the supreme importance of $\mathbf{T}$ in GR. Newton's theory has as a source of the field the density ρ. This was understood to be the mass density, and so is closest to our ρ_0. But a theory that uses rest mass only as its source would be peculiar from a relativistic viewpoint, since rest mass and energy are interconvertible. In fact, we can show that such a theory would violate some very high-precision experiments (to be discussed later). So the source of the field should be *all* energies, the density of total mass energy T^{00}. But to have as the source of the field only one component of a tensor would give a noninvariant theory of gravity: we would need to choose a preferred frame in order to calculate T^{00}. Therefore Einstein guessed that the source of the field ought to be **T**: all stresses and pressures and momenta must also act as sources. Combining this with his insight into curved spaces led him to GR.

The second comment is about pressure, which plays a more fundamental role in GR than in Newtonian theory: first, because it is a source of the field; and, second, because of its appearance in the $(\rho + p)$ term in Eq. (4.54). Consider a dense star, whose strong gravitational field requires a large pressure gradient. How large is measured by the acceleration the fluid element would have, a_i, in the absence of pressure. Given the field, and hence given a_i, the required pressure gradient is just that which would cause the opposite acceleration without gravity:

$$-a_i = \frac{p_{,i}}{\rho + p}.$$

This gives the pressure gradient $p_{,i}$. Since $(\rho + p)$ is greater than ρ, the gradient must be larger in relativity than in Newtonian theory. Moreover, since all components of **T** are sources of the gravitational field, this larger pressure adds to the gravitational field, causing even larger pressures (compared to Newtonian stars) to be required to hold the star up. For stars where $p \ll \rho$ (see below), this doesn't make much difference. But when p becomes comparable to ρ, we find that increasing the pressure is self-defeating: *no* pressure gradient will hold the star up, and gravitational collapse must occur. This description, of course, glosses over much detailed calculation, but it shows that even by studying fluids in SR we can begin to appreciate some of the fundamental changes GR brings to gravitation.

Let us just remind ourselves of the relative sizes of p and ρ. We saw in Exer. 1, § 1.14, that $p \ll \rho$ in ordinary situations. In fact, we only get $p \approx \rho$ for very dense material (neutron star) or material so hot that the particles move at close to the speed of light (a 'relativistic' gas).

4.8 Gauss ’ law

Our final topic on fluids is the integral form of the conservation laws, which are expressed in differential form in Eqs. (4.34) and (4.35). As in three-dimensional vector calculus, the conversion of a volume integral of a divergence into a surface integral is called Gauss’ law. The proof of the theorem is exactly the same as in three dimensions, so we shall not derive it in detail:

$$\int V^{\alpha}{}_{,\alpha}\,\mathrm{d}^4x = \oint V^{\alpha} n_{\alpha}\,\mathrm{d}^3 S, \tag{4.57}$$

where $\tilde{n}$ is the unit-normal one-form discussed in § 4.3, and $\mathrm{d}^3 S$ denotes the three-volume of the three-dimensional hypersurface bounding the four-dimensional volume of integration. The sense of the normal is that it is *outward* pointing, of course, just as in three dimensions. In Fig. 4.9 a simple volume is drawn, in order to illustrate the meaning of Eq. (4.57). The volume is bounded by four pairs of hypersurfaces, for constant t, x, y, and z; only two pairs are shown, since we can only draw two dimensions easily. The normal on the t_2 surface is $\tilde{\mathrm{d}}\,t$. The normal on the t_1 surface is $-\tilde{\mathrm{d}}\,t$, since ‘outward’ is clearly backwards in time. The normal on x_2 is $\tilde{\mathrm{d}}x$, and on x_1 is $-\tilde{\mathrm{d}}x$. So the surface integral in Eq. (4.57) is

$$\begin{aligned}&\int_{t_2} V^0\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z + \int_{t_1} (-V^0)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z\\ &\quad + \int_{x_2} V^x\mathrm{d}t\,\mathrm{d}y\,\mathrm{d}z + \int_{x_1} (-V^x)\,\mathrm{d}t\,\mathrm{d}y\,\mathrm{d}z\\ &\quad + \text{similar terms for the other surfaces in the boundary.}\end{aligned}$$

We can rewrite this as

$$\begin{aligned}&\int \left[V^0(t_2) - V^0(t_1)\right]\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z\\ &\quad + \int \left[V^x(x_2) - V^x(x_1)\right]\mathrm{d}t\,\mathrm{d}y\,\mathrm{d}z + \cdots .\end{aligned} \tag{4.58}$$

If we let $\vec{V}$ be $\vec{N}$, then $N^{\alpha}{}_{,\alpha} = 0$ means that the above expression vanishes, which has the interpretation that change in the number of particles in the three-volume (first integral) is due to the flux across its boundaries (second and subsequent terms). If we are talking about energy conservation, we replace N^{α} with $T^{0\alpha}$, and use $T^{0\alpha}{}_{,\alpha} = 0$. Then, obviously, a similar interpretation of Eq. (4.58) applies. Gauss' law gives an integral version of energy conservation.

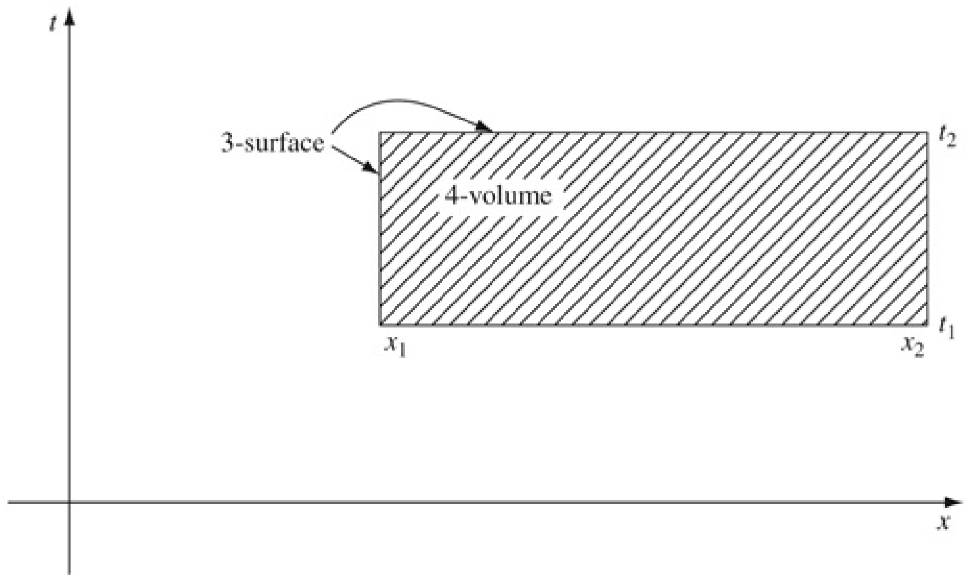


Figure 4.9 The boundary of a region of spacetime.

4.9 Further reading

Continuum mechanics and conservation laws are treated in most texts on GR, such as Misner *et al.* (1973). Students whose background in thermodynamics or fluid mechanics is weak are referred to the classic works of Fermi (1956) and Landau and Lifshitz (1959) respectively. Apart from Exer. 25, § 4.10 below, we do not study much about electromagnetism, but it has a stress–energy tensor and illustrates conservation laws particularly clearly. See Landau and Lifshitz (1962) or Jackson (1975). Relativistic fluids with dissipation present their own difficulties, which reward close study. See Israel and Stewart (1980). Another model for continuum systems is the collisionless gas; see Andréasson (2005) for a description of how to treat such systems in GR.

4.10 Exercises

Comment on whether the continuum approximation is likely to apply to the following physical systems: (a) planetary motions in the solar system; (b) lava flow from a volcano; (c) traffic on a major road at rush hour; (d) traffic at an intersection controlled by stop signs for each incoming road; (e) plasma dynamics.

Flux across a surface of constant x is often loosely called 'flux in the x direction'. Use your understanding of vectors and one-forms to argue that this is an inappropriate way of referring to a flux.

(a) Describe how the Galilean concept of momentum is frame dependent in a manner in which the relativistic concept is not.

(b) How is this possible, since the relativistic definition is nearly the same as the Galilean one for small velocities? (Define a *Galilean* four-momentum vector.)

Show that the number density of dust measured by an arbitrary observer whose four-velocity is $\vec{U}_{\text{obs}}$ is $-\vec{N} \cdot \vec{U}_{\text{obs}}$.

Complete the proof that Eq. (4.14) defines a tensor by arguing that it must be linear in both its arguments.

Establish Eq. (4.19) from the preceding equations.

Derive Eq. (4.21).

) Argue that Eqs. (4.25) and (4.26) can be written as relations among one-forms, *i.e.*

$$\tilde{\mathrm{d}}\rho - (\rho + p)\tilde{\mathrm{d}}n/n = nT\,\tilde{\mathrm{d}}S = n\tilde{\Delta}q.$$

(b) Show that the one-form $\tilde{\Delta}q$ is not a gradient, *i.e.* is not $\tilde{\mathrm{d}}q$ for any function q.

Show that Eq. (4.34), when α is any spatial index, is just Newton's second law.

Take the limit of Eq. (4.35) for $|\mathbf{v}| \ll 1$ to get $\partial n/\partial t + \partial(nv^i)/\partial x^i = 0.$

ı) Show that the matrix δ^{ij} is unchanged when transformed by a rotation of the spatial axes.

(b) Show that any matrix which has this property is a multiple of δ^{ij}.

Derive Eq. (4.37) from Eq. (4.36).

Supply the reasoning in Eq. (4.44).

Argue that Eq. (4.46) is the time component of Eq. (4.45) in the MCRF.

Derive Eq. (4.48) from Eq. (4.47).

In the MCRF, $U^i = 0$. Why can't we assume $U^i{}_{,\beta} = 0$?

We have defined $a^\mu = U^\mu{}_{,\beta}U^\beta$. Go to the nonrelativistic limit (small velocity) and show that $a^i = \dot{v}^i + (\mathbf{v} \cdot \nabla)v^i = Dv^i/Dt,$

where the operator D/Dt is the usual ‘total’ or ‘advective’ time derivative of fluid dynamics.

Sharpen the discussion at the end of § 4.6 by showing that $-\nabla p$ is actually the net force per unit volume on the fluid element in the MCRF.

Show that Eq. (4.58) can be used to prove Gauss’ law, Eq. (4.57).

ı) Show that if particles are not conserved but are generated locally at a rate ε particles per unit volume per unit time in the MCRF, then the conservation law, Eq. (4.35), becomes

$$N^{\alpha}{}_{,\alpha} = \varepsilon.$$

(b) Generalize (a) to show that if the energy and momentum of a body are not conserved (e.g. because it interacts with other systems), then there is a nonzero relativistic force four-vector F^{α} defined by

$$T^{\alpha\beta}{}_{,\beta} = F^{\alpha}.$$

Interpret the components of F^{α} in the MCRF.

In an inertial frame $\mathcal{O}$ calculate the components of the stress–energy tensors of the following systems: (a) A group of particles all moving with the same velocity $\mathbf{v} = \beta \mathbf{e}_x$, as seen in $\mathcal{O}$. Let the rest-mass density of these particles be ρ_0, as measured in their comoving frame. Assume a sufficiently high density of particles to enable treating them as a continuum.

(b) A ring of N similar particles of mass m rotating counter-clockwise in the $x - y$ plane about the origin of $\mathcal{O}$, at a radius a from this point, with an angular velocity ω. The ring is a torus of circular cross-section of radius $\delta a \ll a$, within which the particles are uniformly distributed with a high enough density for the continuum approximation to apply. Do not include the stress–energy of whatever forces keep them in orbit. (Part of the calculation will relate ρ_0 of part (a) to N, a, ω, and δa.) (c) Two such rings of particles, one rotating clockwise and the other counter-clockwise, at the same radius a. The particles do not collide or interact in any way.

Many physical systems may be idealized as collections of noncolliding particles (for example, black-body radiation, rarified plasmas, galaxies, and globular clusters). By assuming that such a system has a random distribution of velocities at every point, with no bias in any direction in the MCRF, prove that the stress–energy tensor is that of a perfect fluid. If all particles have the same speed υ and mass m, express p and ρ as functions of m, υ, and n. Show that a photon gas has

$p = \frac{1}{3}\rho$.

Use the identity $T^{\mu\nu}{}_{,\nu} = 0$ to prove the following results for a bounded system (i.e. a system for which $T^{\mu\nu} = 0$ outside a bounded region of space):

(a) $\frac{\partial}{\partial t}\int T^{0\alpha}\,\mathrm{d}^3x = 0$ (conservation of energy and momentum).

(b) $\frac{\partial^2}{\partial t^2}\int T^{00}x^i x^j\,\mathrm{d}^3x = 2\int T^{ij}\,\mathrm{d}^3x$ (tensor virial theorem).

(c) $\frac{\partial^2}{\partial t^2}\int T^{00}(x^i x_i)^2\,\mathrm{d}^3x = 4\int T^i{}_i x^j x_j\,\mathrm{d}^3x + 8\int T^{ij}x_i x_j\,\mathrm{d}^3x$.

Astronomical observations of the brightness of objects are measurements of the flux of radiation T^{0i} from the object at Earth. This problem calculates how that flux depends on the relative velocity of the object and Earth.

(a) Show that, in the rest frame $\mathcal{O}$ of a star of constant luminosity L (total energy radiated per second), the stress–energy tensor of the radiation from the star at the event $(t, x, 0, 0)$ has components $T^{00} = T^{0x} = T^{x0} = T^{xx} = L/(4\pi x^2)$. The star sits at the origin.

(b) Let $\vec{X}$ be the null vector that separates the events of emission and reception of the radiation. Show that $\vec{X} \to_{\mathcal{O}} (x, x, 0, 0)$ for radiation observed at the event $(x, x, 0, 0)$. Show that the stress–energy tensor of (a) has the frame-invariant form

$$\mathbf{T} = \frac{L}{4\pi}\frac{\vec{X}\otimes\vec{X}}{(\vec{U}_s\cdot\vec{X})^4},$$

where $\vec{U}_s$ is the star's four-velocity, $\vec{U}_s \to_{\mathcal{O}} (1, 0, 0, 0)$.

(c) Let the Earth-bound observer $\bar{\mathcal{O}}$, traveling with speed v away from the star in the x direction, measure the same radiation, again with the star on the $\bar{x}$ axis. Let $\vec{X} \to (R, R, 0, 0)$ and find R as a function of x. Express $T^{\bar{0}\bar{x}}$ in terms of R. Explain why R and $T^{\bar{0}\bar{x}}$ depend as they do on v.

Electromagnetism in SR. (This exercise is suitable only for students who have already encountered Maxwell's equations in some form.) Maxwell's equations for the electric and magnetic fields in vacuum, **E** and **B**, in three-vector notation are

$$\begin{aligned}\nabla\times\mathbf{B} - \frac{\partial}{\partial t}\mathbf{E} &= 4\pi\mathbf{J},\\ \nabla\times\mathbf{E} + \frac{\partial}{\partial t}\mathbf{B} &= 0,\\ \nabla\cdot\mathbf{E} &= 4\pi\rho,\\ \nabla\cdot\mathbf{B} &= 0,\end{aligned} \qquad (4.59)$$

in units where $\mu_0 = \varepsilon_0 = c = 1$. (Here ρ is the density of electric charge and **J** the current density.) (a) An *antisymmetric* (2) tensor **F** can be defined on spacetime by the equations 0 $F^{0i} = E^i (i = 1, 2, 3)$, $F^{xy} = B^z$, $F^{yz} = B^x$, $F^{zx} = B^y$. Find from this definition all other components $F^{\mu\nu}$ in this frame and write them down in a matrix.

(b) A rotation by an angle θ about the z axis is one kind of Lorentz transformation, with the matrix

$$\Lambda^{\beta'}{}_{\alpha} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta & 0 \\ 0 & \sin\theta & \cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Show that the new components of **F**,

$$F^{\alpha'\beta'} = \Lambda^{\alpha'}{}_{\mu}\Lambda^{\beta'}{}_{\nu}F^{\mu\nu},$$

define new electric and magnetic three-vector components (by the rule given in (a)) that are just the same as the components of the old **E** and **B** in the rotated three-space. (This shows that a spatial rotation of **F** makes a spatial rotation of **E** and **B**.) (c) Define the current four-vector $\vec{J}$ by $J^0 = \rho$, $J^i = (\vec{J})^i$, and show that two of Maxwell's equations are just

$$F^{\mu\nu}{}_{,\nu} = 4\pi J^{\mu}. \tag{4.60}$$

(d) Show that the other two of Maxwell's equations are

$$F_{\mu\nu,\lambda} + F_{\nu\lambda,\mu} + F_{\lambda\mu,\nu} = 0. \tag{4.61}$$

Note that there are only *four* independent equations here. That is, choose one index value, say zero. Then the three other values (1, 2, 3) can be assigned to μ, ν, λ in *any* order, producing the same equation (up to an overall sign) each time. Try it and see: it follows from antisymmetry of $F_{\mu\nu}$.

(e) We have now expressed Maxwell's equations in tensor form. Show that conservation of charge, $J^{\mu}{}_{,\mu} = 0$ (recall the similar Eq. (4.35) for the number–flux vector $\vec{N}$), is implied by Eq. (4.60) above. (Hint: use antisymmetry of $F_{\mu\nu}$.) (f) The charge density in any frame is J^0. Therefore the total charge in spacetime is $Q = \int J^0 \, dx \, dy \, dz$, where the integral extends over an entire hypersurface t = const. Defining $\tilde{d}t = \tilde{n}$, a unit normal for this hypersurface, show that

$$Q = \int J^\alpha n_\alpha \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z. \tag{4.62}$$

(g) Use Gauss' law and Eq. (4.60) to show that the total charge enclosed within any closed two-surface $\mathcal{S}$ in the hypersurface t = const. can be determined by doing an integral over $\mathcal{S}$ itself: $Q = \oint_{\mathcal{S}} F^{0i} n_i \, \mathrm{d}\mathcal{S} = \oint_{\mathcal{S}} \mathbf{E} \cdot \mathbf{n} \, \mathrm{d}\mathcal{S},$

where **n** is the unit normal to $\mathcal{S}$ in the hypersurface (*not* the same as $\tilde{n}$ in part (f) above).

(h) Perform a Lorentz transformation on $F^{\mu\nu}$ to a frame $\bar{\mathcal{O}}$ moving with velocity v in the x direction relative to the frame used in (a) above. In this frame define a three-vector $\bar{\mathbf{E}}$ with components $\bar{E}^i = F^{\bar{0}\bar{i}}$, and similarly for $\bar{\mathbf{B}}$ in analogy with (a). In this way discover how **E** and **B** behave under a Lorentz transformation: they get mixed together! Thus, **E** and **B** themselves are not Lorentz invariant, but are merely components of **F**, called the Faraday tensor, which is *the* invariant description of electromagnetic fields in relativity. If you think carefully, you will see that on physical grounds they *cannot* be invariant. In particular, the magnetic field is created by moving charges; but a charge moving in one frame may be at rest in another, so a magnetic field which exists in one frame may not exist in another. What is the same in *all* frames is the Faraday tensor: only its components get transformed.

5

Preface to curvature

5.1 On the relation of gravitation to curvature

Until now we have discussed only SR. In SR, forces have played a background role, and we have never introduced gravitation explicitly as a possible force. One ingredient of SR is the existence of inertial frames that fill all of spacetime: all of spacetime can be described by a single frame, all of whose coordinate points are always at rest relative to the origin, and all of whose clocks run at the same rate relative to the origin's clock. From the fundamental postulates we were led to the idea of the interval Δs^2, which gives an invariant geometrical meaning to certain physical statements. For example, a timelike interval between two events is the time elapsed on a clock which passes through the two events; a spacelike interval is the length of a rod that joins two events in a frame in which they are simultaneous. The mathematical function that calculates the interval is the metric, and so the metric of SR is defined physically by lengths of rods and readings of clocks. This is the power of SR and one reason for the elegance and compactness of tensor notation in it (for instance the replacement of 'number density' and 'flux' by $\vec{N}$). On a piece of paper on which had been plotted all the events and world lines of interest in some coordinate system, it would always be possible to define *any* metric by just giving its components $g_{\alpha\beta}$ as some arbitrarily chosen set of functions of the coordinates. But this arbitrary metric would be useless in doing physical calculations. The usefulness of $\eta_{\alpha\beta}$ is its close relation to experiment, and our derivation of it drew heavily on the experiments.

This closeness to experiment is, of course, a test. Since $\eta_{\alpha\beta}$ makes certain predictions about rods and clocks, we can ask for their verification. In particular, is it *possible* to construct a frame in which the clocks all run at the same rate? This is a crucial question, and we shall show that in a nonuniform gravitational field the answer, experimentally, is no. In this sense, gravitational fields are incompatible with *global* SR: the ability to construct a global inertial frame. We shall see that in small regions of spacetime – regions small enough that nonuniformities of the gravitational forces are too small to measure – we can always construct a 'local' SR frame. In this sense, we shall have to build local SR into a more general theory. The first step is the proof that clocks don't all run at the same rate in a gravitational field.

The gravitational redshift experiment

Let us first imagine performing an idealized experiment, first suggested by Einstein. (i) Let a tower of height h be constructed on the surface of Earth, as in Fig. 5.1. Begin with a particle of rest mass m at the top of the tower. (ii) The particle is dropped and falls freely with acceleration g. It reaches the ground with velocity $v = (2gh)^{1/2}$, so its total energy E, as measured by an experimenter on the ground, is $m + \frac{1}{2}mv^2 + 0(v^4) = m + mgh + 0(v^4)$. (iii) The experimenter on the ground has some magical method of changing all this energy into a single photon of the same energy, which he directs upwards. (Such a process does not violate conservation laws, since Earth absorbs the photon's momentum but not its energy, just as it does for a bouncing rubber ball. The student skeptical of 'magic' should show how the argument proceeds if only a fraction ε of the energy is converted into a photon.) (iv) Upon its arrival at the top of the tower with energy E', the photon is again magically changed into a particle of rest mass $m' = E'$. It must be that $m' = m$; otherwise, perpetual motion could result by the gain in energy obtained by operating such an experiment. So we are led by our abhorrence of the injustice of perpetual motion to *predict* that $E' = m$ or, for the photon,

$$\frac{E'}{E} = \frac{h\nu'}{h\nu} = \frac{m}{m + mgh + 0(v^4)} = 1 - gh + 0(v^4). \tag{5.1}$$

We predict that a photon climbing in Earth's gravitational field will lose energy (not surprisingly) and will consequently be redshifted.

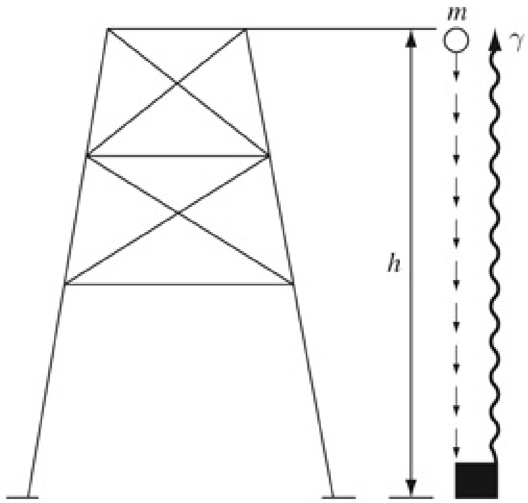


Figure 5.1 A mass m is dropped from a tower of height h. The total mass at the bottom is converted into energy and returned to the top as a photon. Perpetual motion will be performed unless the photon loses as much energy in climbing as the mass gained in falling. Light is therefore redshifted as it climbs in a gravitational field.

Although our thought experiment is too idealized to be practical, it is possible to measure the redshift predicted by Eq. (5.1) directly. This was first done by Pound and Rebka (1960) and improved by Pound and Snider (1965). The experiment used the Mössbauer effect to obtain great precision in the measurement of the difference $\nu' - \nu$ produced in a photon climbing a distance $h = 22.5$ m. Eq. (5.1) was verified to approximately 1% precision.

With improvements in technology between 1960 and 1990, the gravitational redshift moved from being a small exotic correction to becoming an effect that is central to society: the GPS navigation system incorporates vital corrections for the redshift, in the absence of which it would not remain accurate for more than a few minutes. The system uses a network of high-precision atomic clocks in orbiting satellites, and navigation by an apparatus on Earth is accomplished by reading the time-stamps on signals received from five or more satellites. But, as we shall see below, the gravitational redshift implies that time itself runs slightly faster at the higher altitude than it does on the Earth. If this were not compensated for, the ground receiver would soon get wrong time-stamps. The successful operation of GPS can be taken to be a very accurate verification of the redshift. See Ashby (2003) for a full discussion of relativity and the GPS system.

This experimental verification of the redshift is comforting from the point of view of energy conservation. But it is the death-blow to our chances of finding a simple, special-relativistic theory of gravity, as we shall now show.

Nonexistence of a Lorentz frame at rest on Earth

If SR is to be valid in a gravitational field, it is a natural first guess to assume that the 'laboratory' frame at rest on Earth is a Lorentz frame. The following argument, due originally to Schild (1967), easily shows this assumption to be false. In Fig. 5.2 we draw a spacetime diagram in this hypothetical frame, in which the one spatial dimension plotted is the vertical one. Consider light as a wave, and look at two successive 'crests' of the wave as they move upward in the Pound–Rebka–Snider experiment. The top and bottom of the tower have vertical world lines in this diagram, since they are at rest. Light is shown moving on a wiggly line, and it is purposely drawn curved in some arbitrary way. This is to allow for the possibility that gravity may act on light in an unknown way, deflecting it from a null path. But no matter how light is affected by gravity the effect must be the same on both wave crests, since the gravitational field does not change from one time to another. Therefore the two crests' paths are *congruent*, and we conclude from the hypothetical Minkowski geometry that $\Delta t_{top} = \Delta t_{bottom}$. On the other hand, the time between two crests is simply the reciprocal of the measured frequency $\Delta t = 1/\nu$. Since the Pound– Rebka–Snider experiment establishes that $\nu_{bottom} > \nu_{top}$, we know that $\Delta t_{top} > \Delta t_{bottom}$. The conclusion from Minkowski geometry is wrong, and the reference frame at rest on Earth is not a Lorentz frame.

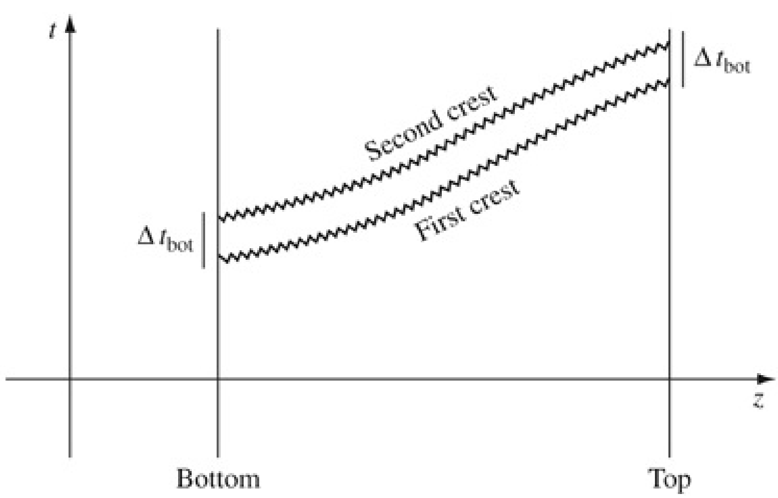


Figure 5.2 In a time-independent gravitational field, two successive 'crests' of an electromagnetic wave must travel identical paths. Because of the redshift (Eq. (5.1)) the time between them at the top is larger than at the bottom. An observer

at the top therefore ‘sees’ a clock at the bottom running slowly.

Is this the end, then, of SR? Not quite. We have shown that the Lorentz frame at rest on Earth is not inertial. We have not shown that there are *no* inertial frames. In fact there are certain frames which are inertial in a restricted sense, and in the next paragraph we shall use another physical argument to find them.

The principle of equivalence

One important property of an inertial frame is that a particle at rest in it stays at rest if no forces act on it. In order to use this, we must have an idea of what a force is. Ordinarily, gravity is regarded as a force. But, as Galileo demonstrated in his famous experiment at the Leaning Tower of Pisa, gravity is distinguished from all other forces in a remarkable way: all bodies given the same initial velocity follow the same trajectory in a gravitational field, regardless of their internal composition. With all other forces, some bodies are affected and others are not: electromagnetism affects charged particles but not neutral ones, and the trajectory of a charged particle depends on the ratio of its charge to its mass, which is not the same for all particles. Similarly, the other two basic forces in physics – the so-called ‘strong’ and ‘weak’ interactions – affect different particles differently. With all these forces, it would always be possible to define experimentally the trajectory of a particle unaffected by the force, *i.e.* a particle that remained at rest in an inertial frame. But, with gravity, this does not work. Attempting to define an inertial frame at rest on Earth, then, is vacuous, since *no* free particle (not even a photon) could possibly be a physical ‘marker’ for it.

But there is a frame in which particles do keep a uniform velocity. This is a frame which falls freely in the gravitational field. Since this frame accelerates at the same rate as free particles do (at least the low-velocity particles to which Newtonian gravitational physics applies), it follows that all such particles will maintain a uniform velocity relative to this frame. This frame is at least a candidate for an inertial frame. In the next section we will show that photons are not redshifted in this frame, which makes it an even better candidate. Einstein built GR by taking the hypothesis that these frames are inertial.

The argument we have just made, that freely falling frames are inertial, will perhaps be more familiar to the student if it is turned around. Consider, in empty space free of gravity, a uniformly accelerating rocket ship. From the point of view of an observer inside, it appears that there is a gravitational field in the rocket: objects dropped accelerate toward the rear of the ship, all with the same acceleration, independent of their internal composition.[1] Moreover, an object held stationary relative to the ship has 'weight' equal to the force required to keep it accelerating with the ship. Just as in 'real' gravity, this force is proportional to the mass of the object. A true inertial frame is one which falls freely toward the rear of the ship, at the same acceleration as particles. From this it can be seen that uniform gravitational fields are equivalent to frames that accelerate uniformly relative to inertial frames. This is the *principle of equivalence* between gravity and acceleration, and is a cornerstone of Einstein's theory. Although Galileo and Newton would have used different words to describe it, the equivalence principle is one of the foundations of Newtonian gravity.

In more modern terminology, what we have described is called the *weak equivalence principle*, 'weak' because it refers to the way bodies behave only when influenced by gravity. Einstein realized that, in order to create a full theory of gravity, he had to extend this to include the other laws of physics. What we now call the *Einstein equivalence principle* says that we can discover how all the other forces of nature behave in a gravitational field by postulating that the differential equations that describe the laws of physics have the same local form in a freely falling inertial frame as they do in SR, *i.e.* when there are no gravitational fields. We shall use this stronger form of the principle of equivalence in Ch. 7.

Before we return to the proof that freely falling frames are inertial, even for photons, we must make two important observations. The first is that our arguments are valid only locally – since the gravitational field of Earth is not uniform, particles some distance away do not remain at uniform velocity in a particular freely falling frame. We shall discuss this in some detail below. The second point is that there are of course an infinity of freely falling frames at any point. They differ in their velocities and in the orientation of their spatial axes, but they all accelerate relative to Earth at the same rate.

The redshift experiment again

Let us now take a different point of view on the Pound–Rebka–Snider experiment. Let us view it in a freely falling frame, which we have seen has at least some of the characteristics of an inertial frame. Let us take the particular frame that is at rest when the photon begins its upward journey and falls freely after that. Since the photon moves a distance h, it takes time $\Delta t = h$ to arrive at the top. In this time, the frame has acquired velocity gh downward relative to the experimental apparatus. So the photon's frequency relative to the freely falling frame can be obtained by the redshift formula

$$\frac{\nu(\text{freely falling})}{\nu'(\text{apparatus at top})} = \frac{1+gh}{\sqrt{(1-g^2h^2)}} = 1 + gh + 0(v^4). \tag{5.2}$$

From Eq. (5.1) we see that if we neglect terms of higher order (as we did to derive Eq. (5.1)), then we get *ν*(photon emitted at bottom) = *ν*(in freely falling frame when photon arrives at top). So there is *no* redshift in a freely falling frame. This gives us a sound basis for postulating that the freely falling frame is an inertial frame.

Local inertial frames

The above discussion suggests that the gravitational redshift experiment really does not render SR and gravity incompatible. Perhaps we simply have to realize that the frame at rest on Earth is not inertial and the freely falling one – in which there is no redshift and so Fig. 5.2 leads to no contradiction – is the true inertial frame. Unfortunately, this doesn't completely save SR, for the simple reason that the freely falling frames on different sides of Earth fall in different directions: there is *no* single global frame which is everywhere freely falling in Earth's gravitational field and which is still rigid, in that the distances between its coordinate points are constant in time. It is still impossible to construct a *global* inertial frame, and so the most we can salvage is a *local* inertial frame, which we now describe.

Consider a freely falling frame in Earth's gravitational field. An inertial frame in SR fills all of spacetime, but this freely falling frame would not be inertial if it were extended too far horizontally, because then it would not be falling vertically. In Fig. 5.3 the frame is freely falling at *B*, but at *A* and *C* the motion is not along the trajectory of a test particle. Moreover, since the acceleration of gravity changes with height, the frame cannot remain inertial if extended over too large a vertical distance; if it were falling with particles at one height, it would not be at another. Finally, the frame can have only a limited extent in time as well, since, as it falls, both the above limitations become more severe due to the frame's approaching closer to Earth. All of these limitations are due to nonuniformities in the gravitational field. Insofar as nonuniformities can be neglected, the freely falling frame can be regarded as inertial. *Any* gravitational field can be regarded as uniform over a small enough region of space and time, and so we can always set up *local* inertial frames. They are analogous to the MCRFs of fluids: in this case the frame is inertial in only a small region for a small time. How small depends on (a) the strength of the nonuniformities of the gravitational field, and (b) the sensitivity of whatever experiment is being used to detect noninertial properties of the frame. Since *any* nonuniformity is, in principle, detectable, a frame can only be regarded mathematically as inertial in a vanishingly small region. But for current technology, the freely falling frames near the surface of Earth can be regarded as inertial to a high accuracy. We will be more quantitative in a later chapter. For now, we just emphasize the mathematical notion that any theory of gravity must admit *local inertial frames*: frames that, at a point, are inertial frames of SR.

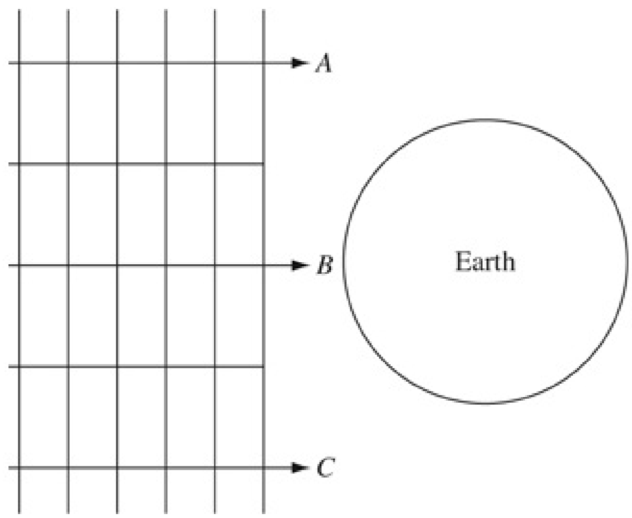


Figure 5.3 A rigid frame cannot fall freely in the Earth’s field and still remain rigid.

Tidal forces

Nonuniformities in gravitational fields are called tidal forces, since they are the ones that raise tides. (If Earth were in a uniform gravitational field, it would fall freely and have no tides. Tides bulge due to the *difference* of the Moon’s and Sun’s gravitational fields across the diameter of Earth.) We have seen that these tidal forces prevent the construction of global inertial frames. It is therefore these forces that are regarded as the fundamental manifestation of gravity in GR.

The role of curvature

The world lines of free particles have been our probe of the possibility of constructing inertial frames. In SR, two such world lines which begin parallel to each other remain parallel, no matter how far they are extended. This is exactly the property that straight lines have in Euclidean geometry. It is natural, therefore, to discuss the *geometry* of spacetime as defined by the world lines of free particles. In these terms, Minkowski space is a *flat* space, because it obeys Euclid's parallelism axiom. It is not a Euclidean space, however, since its metric is different: photons travel on straight world lines of zero proper length. So SR has a flat, nonEuclidean geometry.

Now, in a nonuniform gravitational field, the world lines of two nearby particles which begin parallel do not generally remain parallel. Gravitational spacetime is therefore not flat. In Euclidean geometry, when we drop the parallelism axiom, we get a curved space. For example, the surface of a sphere is curved. Locally straight lines on a sphere extend to great circles, and two great circles always intersect. Nevertheless, sufficiently near to any point, we can pretend that the geometry is flat: the map of a town can be represented on a flat sheet of paper without significant distortion, while a similar attempt for the whole globe fails completely. The sphere is thus locally flat. This is true for all so-called Riemannian[2] spaces: they all are locally flat, but the locally straight lines (called *geodesics*) do not usually remain parallel.

Einstein's important advance was to see the similarity between Riemannian spaces and gravitational physics. He identified the trajectories of freely falling particles with the geodesics of a curved geometry: they are locally straight since spacetime admits local inertial frames in which those trajectories are straight lines, but globally they do not remain parallel.

We shall follow Einstein and look for a theory of gravity that uses a curved spacetime to represent the effects of gravity on particles' trajectories. To do this we shall clearly have to study the mathematics of curvature. The simplest introduction is actually to study curvilinear coordinate systems in a flat space, where our intuition is soundest. We shall see that this will develop nearly all the mathematical concepts we need, and the step to a curved space will be simple. So for the rest of this chapter we will study the Euclidean plane: no more SR (for the time being!) and no more indefinite inner products. What we are after in this chapter is parallelism, not metrics. This approach has the added bonus of giving a more sensible derivation to such often-mysterious formulae as the expression for ∇^2 in polar coordinates!

5.2 Tensor algebra in polar coordinates

Consider the Euclidean plane. The usual coordinates are x and y. Sometimes polar coordinates $\{r, \theta\}$ are convenient:

$$\left.\begin{aligned} r &= (x^2+y^2)^{1/2}, \quad & x &= r\cos\theta, \\ \theta &= \arctan(y/x), \quad & y &= r\sin\theta. \end{aligned}\right\} \tag{5.3}$$

Small increments Δr and $\Delta\theta$ are produced by Δx and Δy according to

$$\left.\begin{aligned} \Delta r &= \tfrac{x}{r}\Delta x + \tfrac{y}{r}\Delta y = \cos\theta\,\Delta x + \sin\theta\,\Delta y, \\ \Delta\theta &= -\tfrac{y}{r^2}\Delta x + \tfrac{x}{r^2}\Delta y = -\tfrac{1}{r}\sin\theta\,\Delta x + \tfrac{1}{r}\cos\theta\,\Delta y, \end{aligned}\right\} \tag{5.4}$$

which are valid to first order.

It is also possible to use other coordinate systems. Let us denote a general coordinate system by ξ and η:

$$\left.\begin{aligned} \xi &= \xi(x,y), \quad & \Delta\xi &= \tfrac{\partial\xi}{\partial x}\Delta x + \tfrac{\partial\xi}{\partial y}\Delta y, \\ \eta &= \eta(x,y), \quad & \Delta\eta &= \tfrac{\partial\eta}{\partial x}\Delta x + \tfrac{\partial\eta}{\partial y}\Delta y. \end{aligned}\right\} \tag{5.5}$$

In order for (ξ, η) to be good coordinates, it is necessary that any two distinct points (x_1, y_1) and (x_2, y_2) be assigned different pairs (ξ_1, η_1) and (ξ_2, η_2), by Eq. (5.5). For instance, the definitions $\xi = x$, $\eta = 1$ would not give good coordinates, since the distinct points $(x = 1, y = 2)$ and $(x = 1, y = 3)$ both have $(\xi = 1, \eta = 1)$. Mathematically, this requires that if $\Delta\xi = \Delta\eta = 0$ in Eq. (5.5), then the points must be the same, or $\Delta x = \Delta y = 0$. This will be true if the determinant of Eq. (5.5) is nonzero,

$$\det\begin{pmatrix} \partial\xi/\partial x & \partial\xi/\partial y \\ \partial\eta/\partial x & \partial\eta/\partial y \end{pmatrix} \neq 0. \tag{5.6}$$

This determinant is called the *Jacobian* of the coordinate transformation, Eq. (5.5). If the Jacobian vanishes at a point, the transformation is said to be *singular* there.

Vectors and one-forms

The old way of defining a vector is to say that it transforms under an *arbitrary* coordinate transformation in the way that the displacement transforms. That is, a vector $\vec{\Delta r}$ can be represented[3] as a displacement $(\Delta x, \Delta y)$, or in polar coordinates $(\Delta r, \Delta\theta)$, or in general $(\Delta\xi, \Delta\eta)$. Then it is clear that for *small* $(\Delta x, \Delta y)$ we have (from Eq. (5.5))

$$\begin{pmatrix}\Delta\xi \\ \\ \Delta\eta\end{pmatrix} = \begin{pmatrix}\partial\xi/\partial x & \partial\xi/\partial y \\ \\ \partial\eta/\partial x & \partial\eta/\partial y\end{pmatrix}\begin{pmatrix}\Delta x \\ \\ \Delta y\end{pmatrix}. \tag{5.7}$$

By defining the matrix of transformation

$$(\Lambda^{\alpha'}{}_{\beta}) = \begin{pmatrix}\partial\xi/\partial x & \partial\xi/\partial y \\ \partial\eta/\partial x & \partial\eta/\partial y\end{pmatrix}, \tag{5.8}$$

we can write the transformation for an arbitrary $\vec{V}$ in the same manner as in SR

$$V^{\alpha'} = \Lambda^{\alpha'}{}_{\beta} V^{\beta}, \tag{5.9}$$

where unprimed indices refer to (x, y) and primed to (ξ, η), and where indices can only take the values 1 and 2. A vector can be defined as something whose components transform according to Eq. (5.9). There is a more sophisticated and natural way, however. This is the modern way, which we now introduce.

Consider a scalar field ϕ. Given coordinates (ξ, η) it is always possible to form the derivatives $\partial\phi/\partial\xi$ and $\partial\phi/\partial\eta$. We *define* the one-form $\tilde{d}\phi$ to be the geometrical object whose components are

$$\tilde{d}\phi \to (\partial\phi/\partial\xi, \partial\phi/\partial\eta) \tag{5.10}$$

in the (ξ, η) coordinate system. This is a general definition of an infinity of one-forms, each formed from a different scalar field. The transformation of components is automatic from the chain rule for partial derivatives:

$$\frac{\partial\phi}{\partial\xi} = \frac{\partial x}{\partial\xi}\frac{\partial\phi}{\partial x} + \frac{\partial y}{\partial\xi}\frac{\partial\phi}{\partial y}, \tag{5.11}$$

and similarly for $\partial\phi/\partial\eta$. The most convenient way to write this in matrix notation is as a transformation on *row-vectors*,

$$(\partial\phi/\partial\xi\ \ \partial\phi/\partial\eta) = (\partial\phi/\partial x\quad \partial\phi/\partial y)\begin{pmatrix}\partial x/\partial\xi & \partial x/\partial\eta \\ \partial y/\partial\xi & \partial y/\partial\eta\end{pmatrix}, \tag{5.12}$$

because then the transformation matrix for one-forms is defined by analogy with Eq. (5.8) as a set of derivatives of the (x, y)-coordinates by the (ξ, η)-coordinates:

$$(\Lambda^{\alpha}{}_{\beta'}) = \begin{pmatrix} \partial x/\partial \xi & \partial x/\partial \eta \\ \partial y/\partial \xi & \partial y/\partial \eta \end{pmatrix}. \tag{5.13}$$

Using this matrix the component-sum version of the transformation in Eq. (5.12) is

$$(\tilde{\mathrm{d}}\phi)_{\beta'} = \Lambda^{\alpha}{}_{\beta'}(\tilde{\mathrm{d}}\phi)_{\alpha}. \tag{5.14}$$

Note that the summation in this equation is on the *first* index of the transformation matrix, as we expect when a row-vector pre-multiplies a matrix.

It is interesting that in SR we did not have to worry about row-vectors, because the simple Lorentz transformation matrices we used were symmetric. But if we want to go beyond even the simplest situations, we need to see that one-form components are elements of row-vectors. However, matrix notation becomes awkward when we go beyond tensors with two indices. In GR we need to deal with tensors with four indices, and sometimes even five. As a result, we will normally express transformation equations in their algebraic form, as in Eq. (5.14); students will not see much matrix notation later in this book.

What we have seen in this section is that, in the modern view, the foundation of tensor algebra is the definition of a one-form. This is more natural than the old way, in which a *single* vector $(\Delta x, \Delta y)$ was defined and others were obtained by analogy. Here a whole *class* of one-forms is defined in terms of derivatives, and the transformation properties of one-forms follow automatically.

Now a vector is defined as a linear function of one-forms into real numbers. The implications of this will be explored in the next paragraph. First we just note that all this is the same as we had in SR, so that vectors do in fact obey the transformation law, Eq. (5.9). It is of interest to see explicitly that $(\Lambda^{\alpha'}{}_{\beta})$ and $(\Lambda^{\alpha}{}_{\beta'})$ are inverses of each other. The product of the matrices is

$$\begin{aligned} &\begin{pmatrix} \partial \xi/\partial x & \partial \xi/\partial y \\ \partial \eta/\partial x & \partial \eta/\partial y \end{pmatrix} \begin{pmatrix} \partial x/\partial \xi & \partial x/\partial \eta \\ \partial y/\partial \xi & \partial y/\partial \eta \end{pmatrix} \\ &= \begin{pmatrix} \frac{\partial \xi}{\partial x}\frac{\partial x}{\partial \xi} + \frac{\partial \xi}{\partial y}\frac{\partial y}{\partial \xi} & \frac{\partial \xi}{\partial x}\frac{\partial x}{\partial \eta} + \frac{\partial \xi}{\partial y}\frac{\partial y}{\partial \eta} \\ \frac{\partial \eta}{\partial x}\frac{\partial x}{\partial \xi} + \frac{\partial \eta}{\partial y}\frac{\partial y}{\partial \xi} & \frac{\partial \eta}{\partial x}\frac{\partial x}{\partial \eta} + \frac{\partial \eta}{\partial y}\frac{\partial y}{\partial \eta} \end{pmatrix}. \end{aligned} \tag{5.15}$$

By the chain rule this matrix is

$$\begin{pmatrix} \partial \xi/\partial \xi & \partial \xi/\partial \eta \\ \partial \eta/\partial \xi & \partial \eta/\partial \eta \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{5.16}$$

where the equality follows from the definition of a partial derivative.

Curves and vectors

Curves and vectors

The usual notion of a curve is of a connected series of points in the plane. This we shall call a *path*, and reserve the word curve for a parametrized path. That is, we shall follow modern mathematical terminology and define a *curve* as a mapping of an interval of the real line into a path in the plane. What this means is that a curve is a path with a real number associated with each point on the path. This number is called the parameter s. Each point has coordinates which may then be expressed as a function of s:

$$\text{Curve} :\{\xi = f(s),\ \eta = g(s), a \leqslant s \leqslant b\} \tag{5.17}$$

defines a curve in the plane. If we were to change the parameter (but not the points) to $s' = s'(s)$, which is a function of the old s, then we would have

$$\{\xi = f'(s'),\ \eta = g'(s'),\ a' \leqslant s' \leqslant b'\}, \tag{5.18}$$

where f' and g' are *new* functions, and where $a' = s'(a)$, $b' = s'(b)$. This is, mathematically, a *new* curve, even though its *image* (the points of the plane that it passes through) is the same. So there is an infinite number of curves having the same path.

The derivative of a scalar field ϕ along the curve is $d\phi/ds$. This depends on s, so by changing the parameter we change the derivative. We can write this as

$$d\phi/ds = \langle \tilde{d}\phi, \vec{V} \rangle, \tag{5.19}$$

where $\vec{V}$ is the vector whose components are $(d\xi/ds, d\eta/ds)$. This vector depends only on the curve, while $\tilde{d}\phi$ depends only on ϕ. Therefore $\vec{V}$ is a vector characteristic of the curve, called the *tangent* vector. (It clearly lies tangent to curve: see Fig. 5.4.) So a vector may be regarded as a thing which produces $d\phi/ds$, given ϕ. This leads to the most modern view, that the tangent vector to the curve should be *called* d/ds. Some relativity texts occasionally use this notation. For our purposes, however, we shall just let $\vec{V}$ be the tangent vector whose components are $(d\xi/ds, d\eta/ds)$. Notice that a path in the plane has, at any point, an infinity of tangents, all of them parallel but differing in length. These are to be regarded as vectors tangent to *different* curves, curves that have different parametrizations in a neighborhood of that point. A curve has a *unique* tangent, since the path *and* parameter are given. Moreover, even curves that have identical tangents at a point may not be identical elsewhere. From the Taylor expansion $\xi(s + 1) \approx \xi(s) + d\xi/ds$, we see that $\vec{V}(s)$ stretches approximately from s to $s + 1$ along the curve.

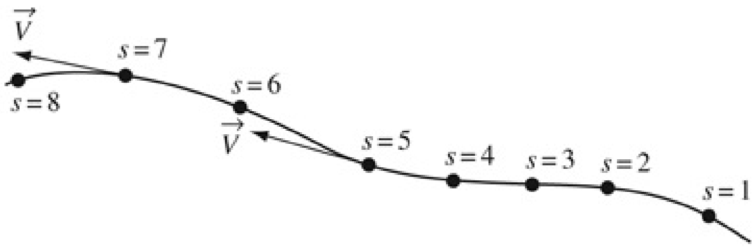


Figure 5.4 A curve, its parametrization, and its tangent vector.

Now, it is clear that under a coordinate transformation s does not change (its definition had nothing to do with coordinates) but the components of $\vec{V}$ will, since by the chain rule

$$\begin{pmatrix} \mathrm{d}\xi/\mathrm{d}s \\ \mathrm{d}\eta/\mathrm{d}s \end{pmatrix} = \begin{pmatrix} \partial\xi/\partial x & \partial\xi/\partial y \\ \partial\eta/\partial x & \partial\eta/\partial y \end{pmatrix} \begin{pmatrix} \mathrm{d}x/\mathrm{d}s \\ \mathrm{d}y/\mathrm{d}s \end{pmatrix}. \tag{5.20}$$

This is the same transformation law as we had for vectors earlier, Eq. (5.7).

To sum up, the modern view is that a vector is a *tangent* to some curve, and is the function that gives $\mathrm{d}\phi/\mathrm{d}s$ when it takes $\tilde{\mathrm{d}}\phi$ as an argument. Having said this, we are now in a position to do polar coordinates more thoroughly.

Polar coordinate basis one-forms and vectors

The bases of the coordinates are clearly

$$\vec{e}_{\alpha'} = \Lambda^{\beta}{}_{\alpha'}\vec{e}_{\beta},$$

or

$$\vec{e}_r = \Lambda^{x}{}_{r}\vec{e}_x + \Lambda^{y}{}_{r}\vec{e}_y \tag{5.21}$$

$$= \frac{\partial x}{\partial r}\vec{e}_x + \frac{\partial y}{\partial r}\vec{e}_y$$

$$= \cos\theta\ \vec{e}_x + \sin\theta\ \vec{e}_y, \tag{5.22}$$

and, similarly,

$$\begin{aligned} \vec{e}_\theta &= \frac{\partial x}{\partial \theta}\vec{e}_x + \frac{\partial y}{\partial \theta}\vec{e}_y \\ &= -r\sin\theta\ \vec{e}_x + r\cos\theta\vec{e}_y. \end{aligned} \tag{5.23}$$

Notice in this that we have used, among others,

$$\Lambda^x{}_r = \frac{\partial x}{\partial r}. \tag{5.24}$$

Similarly, to transform the other way we would need

$$\Lambda^r{}_x = \frac{\partial r}{\partial x}. \tag{5.25}$$

The transformation matrices are exceedingly simple: just keeping track of which index is up and which is down gives the right derivative to use.

The basis one-forms are, analogously,

$$\begin{aligned} \tilde{d}\theta &= \frac{\partial \theta}{\partial x}\,\tilde{d}x + \frac{\partial \theta}{\partial y}\,\tilde{d}y, \\ &= -\frac{1}{r}\sin\theta\,\tilde{d}x + \frac{1}{r}\cos\theta\,\tilde{d}y. \end{aligned} \tag{5.26}$$

(Notice the similarity to ordinary calculus, Eq. (5.4).) Similarly, we find

$$\tilde{d}r = \cos\theta\,\tilde{d}x + \sin\theta\,\tilde{d}y. \tag{5.27}$$

We can draw pictures of the bases at various points (Fig. 5.5). Drawing the basis vectors is no problem. Drawing the basis one-forms is most easily done by drawing surfaces of constant r and θ for $\tilde{d}r$ and $\tilde{d}\theta$. These surfaces have different orientations in different places.

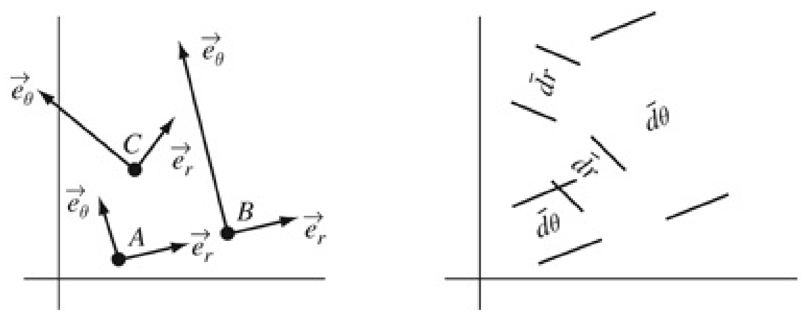


Figure 5.5 Basis vectors and one-forms for polar coordinates.

There is a point of great importance to note here: the bases change from point to point. For the vectors, the basis vectors at A in Fig. 5.5 are not parallel to those at C. This is because they point in the direction of increasing coordinate, which changes from point to point. Moreover, the lengths of the bases are not constant. For example, from Eq. (5.23) we find

$$|\vec{e}_\theta|^2 == \vec{e}_\theta \cdot \vec{e}_\theta = r^2 \sin^2\theta + r^2\cos^2\theta = r^2, \tag{5.28a}$$

so that $\vec{e}_\theta$ increases in magnitude as we get further from the origin. The reason is that the basis vector $\vec{e}_\theta$, having components (0,1) with respect to r and θ, has essentially a θ displacement of one unit, *i.e.* one radian. It must be longer to do this at large radii than at small. So we do not have a *unit* basis. It is easy to verify that

$$|\vec{e}_r| = 1, \quad |\tilde{\mathrm{d}}r| = 1, \quad |\tilde{\mathrm{d}}\theta| = r^{-1}. \tag{5.28b}$$

Again, $|\tilde{\mathrm{d}}\theta|$ gets larger (more intense) near $r = 0$ because a given vector can span a larger range of θ near the origin than farther away.

Metric tensor

The dot products above were all calculated by knowing the metric in Cartesian coordinates x, y:

$$\vec{e}_x \cdot \vec{e}_x = \vec{e}_y \cdot \vec{e}_y = 1, \quad \vec{e}_x \cdot \vec{e}_y = 0;$$

or, put in tensor notation,

$$\mathbf{g}(\vec{e}_\alpha, \vec{e}_\beta) = \delta_{\alpha\beta} \quad \text{in Cartesian coordinates.} \tag{5.29}$$

What are the components of **g** in polar coordinates? Simply

$$g_{\alpha'\beta'} = \mathbf{g}(\vec{e}_{\alpha'}, \vec{e}_{\beta'}) = \vec{e}_{\alpha'} \cdot \vec{e}_{\beta'} \tag{5.30}$$

or, by Eq. (5.28),

$$g_{rr} = 1, \quad g_{\theta\theta} = r^2, \tag{5.31a}$$

and, from Eqs. (5.22) and (5.23),

$$g_{r\theta} = 0. \tag{5.31b}$$

So we can write the components of **g** as

$$(g_{\alpha\beta})_{polar} = \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix}, \tag{5.32}$$

A convenient way of displaying the components of **g** and at the same time showing the coordinates is the line element, which is the magnitude of an arbitrary 'infinitesimal' displacement $\mathrm{d}\vec{l}$:

$$\begin{aligned} \mathrm{d}\vec{l} \cdot \mathrm{d}\vec{l} = \mathrm{d}s^2 &= |\mathrm{d}r\, \vec{e}_r + \mathrm{d}\theta \vec{e}_\theta|^2 \\ &= \mathrm{d}r^2 + r^2 \mathrm{d}\theta^2. \end{aligned} \tag{5.33}$$

Do *not* confuse dr and dθ here with the basis one-forms $\tilde{\mathrm{d}}r$ and $\tilde{\mathrm{d}}\theta$. The things in this equation are components of $\mathrm{d}\vec{l}$ in polar coordinates, and 'd' simply means 'infinitesimal Δ'.

There is another way of deriving Eq. (5.33) which is instructive. Recall Eq. (3.26) in which a general $\binom{0}{2}$ tensor is written as a sum over basis $\binom{0}{2}$ tensors $\tilde{\mathrm{d}}\, x^\alpha \otimes \tilde{\mathrm{d}}x^\beta$. For the metric this is $\mathbf{g} = g_{\alpha\beta}\, \tilde{\mathrm{d}}x^\alpha \otimes \tilde{\mathrm{d}}x^\beta = \tilde{\mathrm{d}}r \otimes \tilde{\mathrm{d}}r + r^2 \tilde{\mathrm{d}}\theta \otimes \tilde{\mathrm{d}}\theta.$ Although this has a superficial resemblance to Eq. (5.33), it is different: it is an operator which, when supplied with the vector $\mathrm{d}\vec{l}$, the components of which are dr and dθ, gives Eq. (5.33). Unfortunately, the two expressions resemble each other rather too closely because of the confusing way notation has evolved in this subject. Most texts and research papers still use the 'old-fashioned' form in Eq. (5.33) for displaying the components of the metric, and we follow the same practice.

The metric has an inverse:

$$\begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & r^{-2} \end{pmatrix}. \tag{5.34}$$

So we have $g^{rr} = 1$, $g^{r\theta} = 0$, $g^{\theta\theta} = 1/r^2$. This enables us to make the mapping between one-forms and vectors. For instance, if ϕ is a scalar field and $\tilde{\mathrm{d}}\phi$ is its gradient, then the vector $\vec{\mathrm{d}}\phi$ has components

$$(\vec{\mathrm{d}}\phi)^\alpha = g^{\alpha\beta}\phi_{,\beta}, \tag{5.35}$$

or

$$\begin{aligned} (\vec{\mathrm{d}}\phi)^r &= g^{r\beta}\phi_{,\beta} = g^{rr}\phi_{,r} + g^{r\theta}\phi_{,\theta} \\ &= \partial\phi/\partial r; \end{aligned} \tag{5.36a}$$

$$\begin{aligned} (\vec{\mathrm{d}}\phi)^\theta &= g^{\theta r}\phi_{,r} + g^{\theta\theta}\phi_{,\theta} \\ &= \frac{1}{r^2}\frac{\partial\phi}{\partial\theta}. \end{aligned} \tag{5.36b}$$

So, while $(\phi_{,r}, \phi_{,\theta})$ are components of a one-form, the vector gradient has components $(\phi_{,r}, \phi_{,\theta}/r^2)$. Even though we are in Euclidean space, vectors generally have different components from their associated one-forms. Cartesian coordinates are the only coordinates in which the components are the same.

5.3 Tensor calculus in polar coordinates

The fact that the basis vectors of polar coordinates are not constant everywhere, leads to some problems when we try to differentiate vectors. For instance, consider the simple vector $\vec{e}_x$, which is a constant vector field, the same at any point. In polar coordinates it has components $\vec{e}_x \to (\Lambda^r{}_x, \Lambda^\theta{}_x) = (\cos\theta, -r^{-1}\sin\theta)$. These are clearly not constant, even though $\vec{e}_x$ is. The reason is that they are components on a nonconstant basis. If we were to differentiate them with respect to, say, θ, we would most certainly *not* get $\partial\vec{e}_x/\partial\theta$, which must be identically zero. So, from this example, we see that differentiating the components of a vector does not necessarily give the derivative of the vector. We must also differentiate the nonconstant basis vectors. This is the key to the understanding of curved coordinates and, indeed, of curved spaces. We shall now make these ideas systematic.

Derivatives of basis vectors

Since $\vec{e}_x$ and $\vec{e}_y$ are constant vector fields, we easily find that

$$\frac{\partial}{\partial r}\vec{e}_r = \frac{\partial}{\partial r}(\cos\theta\,\vec{e}_x + \sin\theta\,\vec{e}_y) = 0, \tag{5.37a}$$

$$\frac{\partial}{\partial\theta}\vec{e}_r = \frac{\partial}{\partial\theta}(\cos\theta\vec{e}_x + \sin\theta\,\vec{e}_y)$$

$$= -\sin\theta\,\vec{e}_x + \cos\theta\vec{e}_y = \frac{1}{r}\,\vec{e}_\theta. \tag{5.37b}$$

These have a simple geometrical picture, shown in Fig. 5.6. At two nearby points, A and B, $\vec{e}_r$ must point directly away from the origin, and so in slightly different directions. The derivative of $\vec{e}_r$ with respect to θ is just the difference between $\vec{e}_r$ at A and B divided by $\Delta\theta$. The difference in this case is clearly a vector parallel to $\vec{e}_\theta$, which then makes Eq. (5.37b) reasonable.

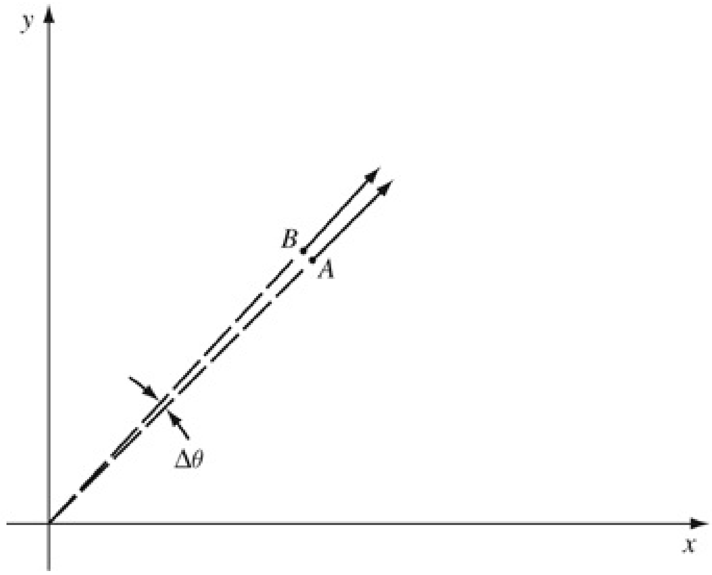


Figure 5.6 Change in $\vec{e}_r$, when $\boldsymbol{\theta}$ changes by $\Delta\theta$.

Similarly,

$$\frac{\partial}{\partial r}\,\vec{e}_\theta = \frac{\partial}{\partial r}\,(-r\sin\theta\,\vec{e}_x + r\cos\theta\,\vec{e}_y)$$
$$= -\sin\theta\vec{e}_x + \cos\theta\,\vec{e}_y = \frac{1}{r}\,\vec{e}_\theta, \tag{5.38a}$$
$$\frac{\partial}{\partial\theta}\,\vec{e}_\theta = -r\cos\theta\,\vec{e}_x - r\sin\,\theta\,\vec{e}_y = -r\,\vec{e}_r. \tag{5.38b}$$

The student is encouraged to draw a picture similar to Fig. 5.6 to explain these formulas.

Derivatives of general vectors

Let us go back to the derivative of $\vec{e}_x$. Since

$$\vec{e}_x = \cos\theta\,\vec{e}_r - \frac{1}{r}\,\sin\theta\,\vec{e}_\theta, \tag{5.39}$$

we have

$$\frac{\partial}{\partial\theta}\,\vec{e}_x = \frac{\partial}{\partial\theta}\,(\cos\theta)\,\vec{e}_r + \cos\theta\frac{\partial}{\partial\theta}\,(\vec{e}_r)$$
$$- \frac{\partial}{\partial\theta}\left(\frac{1}{r}\sin\theta\right)\,\vec{e}_\theta - \frac{1}{r}\sin\theta\frac{\partial}{\partial\theta}\,(\vec{e}_\theta) \tag{5.40}$$
$$= -\sin\theta\,\vec{e}_r + \cos\theta\left(\frac{1}{r}\,\vec{e}_\theta\right)$$
$$- \frac{1}{r}\cos\theta\,\vec{e}_\theta - \frac{1}{r}\sin\theta(-r\,\vec{e}_r). \tag{5.41}$$

To get this we used Eqs. (5.37) and (5.38). Simplifying gives

$$\frac{\partial}{\partial\theta}\,\vec{e}_x = 0, \tag{5.42}$$

just as we should have. Now, in Eq. (5.40) the first and third terms come from differentiating the *components* of $\vec{e}_x$ on the polar coordinate basis; the other two terms are the derivatives of the polar basis vectors themselves, and are necessary for cancelling out the derivatives of the components.

A general vector $\vec{V}$ has components (V^r, V^θ) on the polar basis. Its derivative, by analogy with Eq. (5.40), is

$$\frac{\partial\vec{V}}{\partial r} = \frac{\partial}{\partial r}(V^r\,\vec{e}_r + V^\theta\,\vec{e}_\theta)$$
$$= \frac{\partial V^r}{\partial r}\,\vec{e}_r + V^r\frac{\partial\vec{e}_r}{\partial r} + \frac{\partial V^\theta}{\partial r}\,\vec{e}_\theta + V^\theta\frac{\partial\vec{e}_\theta}{\partial r},$$

and similarly for $\partial\vec{V}/\partial\theta$. Written in index notation, this becomes $\frac{\partial\vec{V}}{\partial r} = \frac{\partial}{\partial r}(V^\alpha\,\vec{e}_\alpha) = \frac{\partial V^\alpha}{\partial r}\,\vec{e}_\alpha + V^\alpha\frac{\partial\vec{e}_\alpha}{\partial r}$.
(Here α runs of course over r and θ.)

This shows explicitly that the derivative of $\vec{V}$ is more than just the derivative of its components V^α. Now, since r is just one coordinate, we can generalize the above equation to

$$\frac{\partial\vec{V}}{\partial x^\beta} = \frac{\partial V^\alpha}{\partial x^\beta}\,\vec{e}_\alpha + V^\alpha\frac{\partial\vec{e}_\alpha}{\partial x^\beta}, \tag{5.43}$$

where, now, x^β can be either r or θ for $\beta = 1$ or 2.

The Christoffel symbols

The final term in Eq. (5.43) is obviously of great importance. Since $\partial\vec{e}_\alpha/\partial x^\beta$ is itself a vector, it can be written as a linear combination of the basis vectors; we introduce the symbol $\Gamma^\mu{}_{\alpha\beta}$ to denote the coefficients in this combination:

$$\frac{\partial \vec{e}_\alpha}{\partial x^\beta} = \Gamma^\mu{}_{\alpha\beta}\, \vec{e}_\mu. \tag{5.44}$$

The interpretation of $\Gamma^\mu{}_{\alpha\beta}$ is that it is the μth component of $\partial \vec{e}_\alpha / \partial x^\beta$. It needs three indices: one (α) gives the basis vector being differentiated; the second (β) gives the coordinate with respect to which it is being differentiated; and the third (μ) denotes the component of the resulting derivative vector. These things, $\Gamma^\mu{}_{\alpha\beta}$, are so useful that they have been given a name: the Christoffel symbols. The question of whether or not they are components of tensors we postpone until much later.

We have of course already calculated them for polar coordinates. From Eqs. (5.37) and (5.38) we find

$$\left.\begin{array}{ll}
(1) & \partial \vec{e}_r/\partial r = 0 \Rightarrow \Gamma^\mu{}_{rr} = 0 \quad \text{for all } \mu, \\
(2) & \partial \vec{e}_r/\partial \theta = \frac{1}{r}\, \vec{e}_\theta \Rightarrow \Gamma^r{}_{r\theta} = 0, \quad \Gamma^\theta{}_{r\theta} = \frac{1}{r}, \\
(3) & \partial \vec{e}_\theta/\partial r = \frac{1}{r}\, \vec{e}_\theta \Rightarrow \Gamma^r{}_{\theta r} = 0, \quad \Gamma^\theta{}_{\theta r} = \frac{1}{r}, \\
(4) & \partial \vec{e}_\theta/\partial \theta = -r\, \vec{e}_r \Rightarrow \Gamma^r{}_{\theta\theta} = -r, \quad \Gamma^\theta{}_{\theta\theta} = 0.
\end{array}\right\} \tag{5.45}$$

In the definition, Eq. (5.44), all indices must refer to the same coordinate system. Thus, although we computed the derivatives of $\vec{e}_r$ and $\vec{e}_\theta$ by using the constancy of $\vec{e}_x$ and $\vec{e}_y$, the Cartesian bases do not in the end make any appearance in Eq. (5.45). The Christoffel symbols' importance is that they enable us to express these derivatives without using any other coordinates than polar.

The covariant derivative

Using the definition of the Christoffel symbols, Eq. (5.44), the derivative in Eq. (5.43) becomes

$$\frac{\partial \vec{V}}{\partial x^\beta} = \frac{\partial V^\alpha}{\partial x^\beta}\, \vec{e}_\alpha + V^\alpha \Gamma^\mu{}_{\alpha\beta}\, \vec{e}_\mu. \tag{5.46}$$

In the last term there are two sums, on α and μ. Relabeling the dummy indices will help here: we change μ to α and α to μ and get

$$\frac{\partial \vec{V}}{\partial x^\beta} = \frac{\partial V^\alpha}{\partial x^\beta}\, \vec{e}_\alpha + V^\mu \Gamma^\alpha{}_{\mu\beta}\, \vec{e}_\alpha. \tag{5.47}$$

The reason for the relabeling was that, now, $\vec{e}_\alpha$ can be factored out of both terms:

$$\frac{\partial \vec{V}}{\partial x^\beta} = \left(\frac{\partial V^\alpha}{\partial x^\beta} + V^\mu \Gamma^\alpha{}_{\mu\beta} \right) \vec{e}_\alpha. \tag{5.48}$$

So the vector field $\partial \vec{V}/\partial x^{\beta}$ has components

$$\partial V^{\alpha}/\partial x^{\beta} + V^{\mu}\Gamma^{\alpha}{}_{\mu\beta}. \tag{5.49}$$

Recall our original notation for the partial derivative, $\partial V^{\alpha}/\partial x^{\beta} = V^{\alpha}{}_{,\beta}$. We keep this notation and define a *new* symbol:

$$V^{\alpha}{}_{;\beta} := V^{\alpha}{}_{,\beta} + V^{\mu}\Gamma^{\alpha}{}_{\mu\beta}. \tag{5.50}$$

Then, with this shorthand semicolon notation, we have

$$\partial \vec{V}/\partial x^{\beta} = V^{\alpha}{}_{;\beta}\, \vec{e}_{\alpha}, \tag{5.51}$$

a very compact way of writing Eq. (5.48).

Now $\partial \vec{V}/\partial x^{\beta}$ is a vector field if we regard β as a given fixed number. But there are two values that β *can* have, and so we can also regard $\partial \vec{V}/\partial x^{\beta}$ as being associated with a $\binom{1}{1}$ tensor field which maps the vector $\vec{e}_{\beta}$ into the vector $\partial\vec{V}/\partial x^{\beta}$, as in Exer. 17, § 3.10. This tensor field is called the *covariant derivative* of $\vec{V}$, denoted, naturally enough, as $\nabla\vec{V}$. Then its components are

$$(\nabla\vec{V})^{\alpha}{}_{\beta} = (\nabla_{\beta}\vec{V})^{\alpha} = V^{\alpha}{}_{;\beta}. \tag{5.52}$$

On a Cartesian basis the components are just $V^{\alpha}{}_{,\beta}$. On the curvilinear basis, however, the derivatives of the basis vectors must be taken into account, and we get that $V^{\alpha}{}_{;\beta}$ are the components of $\nabla\vec{V}$ in whatever coordinate system the Christoffel symbols in Eq. (5.50) refer to. The significance of this statement should not be underrated, as it is the foundation of all our later work. There is a single $\binom{1}{1}$ tensor called $\nabla\vec{V}$. In Cartesian coordinates its components are $\partial V^{\alpha}/\partial x^{\beta}$. In general coordinates $\{x^{\mu'}\}$ its components are called $V^{\alpha'}{}_{;\beta'}$ and can be obtained in either of two equivalent ways: (i) compute them directly in $\{x^{\mu'}\}$ using Eq. (5.50) and a knowledge of what the $\Gamma^{\alpha'}{}_{\mu'\beta'}$ coefficients are in these coordinates; or (ii) obtain them by the usual tensor transformation laws from Cartesian to $\{x^{\mu'}\}$.

What is the covariant derivative of a scalar? The covariant derivative differs from the partial derivative with respect to the coordinates only because the basis vectors change. But a scalar does not depend on the basis vectors, so its covariant derivative is the same as its partial derivative, which is its gradient:

$$\nabla_\alpha f = \partial f/\partial x^\alpha; \quad \nabla f = \tilde{d}f. \tag{5.53}$$

Divergence and Laplacian

Before doing any more theory, let us link this up with things we have seen before. In Cartesian coordinates the divergence of a vector V^α is $V^\alpha{}_{,\alpha}$. This is the scalar obtained by contracting $V^\alpha{}_{,\beta}$ on its two indices. Since contraction is a frame-invariant operation, the divergence of $\vec{V}$ can be calculated in other coordinates $\{x^{\mu'}\}$ also by contracting the components of $\nabla\vec{V}$ on their two indices. This results in a scalar with the value $V^{\alpha'}{}_{;\alpha'}$. It is important to realize that this is the *same* number as $V^\alpha{}_{,\alpha}$ in Cartesian coordinates:

$$V^\alpha{}_{,\alpha} \equiv V^{\beta'}{}_{;\beta'}, \tag{5.54}$$

where unprimed indices refer to Cartesian coordinates and primed refer to the arbitrary system.

For polar coordinates (dropping primes for convenience here)

$$V^\alpha{}_{;\alpha} = \frac{\partial V^\alpha}{\partial x^\alpha} + \Gamma^\alpha{}_{\mu\alpha} V\mu.$$

Now, from Eq. (5.45) we can calculate

$$\left.\begin{aligned} \Gamma^\alpha{}_{r\alpha} &= \Gamma^r{}_{rr} + \Gamma^\theta{}_{r\theta} = 1/r, \\ \Gamma^\alpha{}_{\theta\alpha} &= \Gamma^r{}_{\theta r} + \Gamma^\theta{}_{\theta\theta} = 0. \end{aligned}\right\} \tag{5.55}$$

Therefore we have

$$\begin{aligned} V^\alpha{}_{;\alpha} &= \frac{\partial V^r}{\partial r} + \frac{\partial V^\theta}{\partial \theta} + \frac{1}{r} V^r, \\ &= \frac{1}{r}\frac{\partial}{\partial r}(rV^r) + \frac{\partial}{\partial\theta} V^\theta. \end{aligned} \tag{5.56}$$

This may be a familiar formula to the student. What is probably more familiar is the Laplacian, which is the divergence of the gradient. But we only have the divergence of vectors, and the gradient is a one-form. Therefore we must first convert the one-form to a vector. Thus, given a scalar ϕ, we have the vector gradient (see Eq. (5.53) and the last part of § 5.2 above) with components $(\phi_{,r}, \phi_{,\theta}/r^2)$. Using these as the components of the vector in the divergence formula, Eq. (5.56) gives

$$\nabla \cdot \nabla\phi := \nabla^2\phi = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial\phi}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2\phi}{\partial\theta^2}. \tag{5.57}$$

This is the Laplacian in plane polar coordinates. It is, of course, identically equal to

$$\nabla^2\phi = \frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2}. \tag{5.58}$$

Derivatives of one-forms and tensors of higher types

Since a scalar ϕ depends on no basis vectors, its derivative $\tilde{d}\phi$ is the same as its covariant derivative $\nabla\phi$. We shall almost always use the symbol $\nabla\phi$. To compute the derivative of a one-form (which as for a vector won't be simply the derivatives of its components), we use the property that a one-form and a vector give a scalar. Thus, if $\tilde{p}$ is a one-form and $\vec{V}$ is an arbitrary vector, then for fixed β, $\nabla_\beta\,\tilde{p}$ is also a one-form, $\nabla_\beta\,\vec{V}$ is a vector, and $\langle\tilde{p}, \vec{V}\rangle \equiv \phi$ is a scalar. In any (arbitrary) coordinate system this scalar is just

$$\phi = p_\alpha V^\alpha. \tag{5.59}$$

Therefore $\nabla_\beta\phi$ is, by the product rule for derivatives,

$$\nabla_\beta\phi = \phi_{,\beta} = \frac{\partial p_\alpha}{\partial x^\beta}\,V^\alpha + p_\alpha\frac{\partial V^\alpha}{\partial x^\beta}. \tag{5.60}$$

But we can use Eq. (5.50) to replace $\partial V^\alpha/\partial x^\beta$ in favor of $V^\alpha{}_{;\beta}$, which are the components of $\nabla_\beta\vec{V}$:

$$\nabla_\beta\phi = \frac{\partial p_\alpha}{\partial x^\beta}\,V^\alpha + p_\alpha\,V^\alpha{}_{;\beta} - p_\alpha V^\mu\Gamma^\alpha{}_{\mu\beta}. \tag{5.61}$$

Rearranging terms, and relabeling dummy indices in the term that contains the Christoffel symbol, gives

$$\nabla_\beta\phi = \left(\frac{\partial p_\alpha}{\partial x^\beta} - p_\mu\Gamma^\mu{}_{\alpha\beta}\right)\,V^\alpha + p_\alpha V^\alpha{}_{;\beta}. \tag{5.62}$$

Now, every term in this equation except the one in parentheses is *known* to be the component of a tensor, for an arbitrary vector field $\vec{V}$. Therefore, since multiplication and addition of components always gives new tensors, it must be true that the term in parentheses is also the component of a tensor. This is, of course, the covariant derivative of $\tilde{p}$:

$$(\nabla_\beta \tilde{p})_\alpha := (\nabla \tilde{p})_{\alpha\beta} := p_{\alpha;\beta} = p_{\alpha,\beta} - p_\mu \Gamma^\mu{}_{\alpha\beta}. \tag{5.63}$$

Then Eq. (5.62) reads

$$\nabla_\beta (p_\alpha V^\alpha) = p_{\alpha;\beta} V^\alpha + p_\alpha V^\alpha{}_{;\beta}. \tag{5.64}$$

Thus covariant differentiation obeys the same sort of product rule as Eq. (5.60). It *must* do this, since in Cartesian coordinates ∇ is just partial differentiation of components, so Eq. (5.64) reduces to Eq. (5.60).

Let us compare the two formulae we have, Eq. (5.50) and Eq. (5.63):

$$V^\alpha{}_{;\beta} = V^\alpha{}_{,\beta} + V^\mu \Gamma^\alpha{}_{\mu\beta},$$

$$p_{\alpha;\beta} = p_{\alpha,\beta} - p_\mu \Gamma^\mu{}_{\alpha\beta}.$$

There are certain similarities and certain differences. If we remember that the derivative index β is the *last* one on Γ, then the other indices are the only ones they can be without raising and lowering with the metric. The only thing to watch is the sign difference. It may help to remember that $\Gamma^\alpha{}_{\mu\beta}$ was related to derivatives of the basis vectors, for then it is reasonable that $-\Gamma^\mu{}_{\alpha\beta}$ be related to derivatives of the basis one-forms. The change in sign means that the basis one-forms change 'oppositely' to basis vectors, which makes sense when we remember that the contraction $\langle \tilde{\omega}^\alpha, \vec{e}_\beta \rangle = \delta^\alpha{}_\beta$ is a *constant*, and its derivative must be zero.

The same procedure that led to Eq. (5.63) would lead to the following:

$$\nabla_\beta T_{\mu\nu} = T_{\mu\nu,\beta} - T_{\alpha\nu} \Gamma^\alpha{}_{\mu\beta} - T_{\mu\alpha} \Gamma^\alpha{}_{\nu\beta}; \tag{5.64}$$

$$\nabla_\beta A^{\mu\nu} = A^{\mu\nu}{}_{,\beta} + A^{\alpha\nu} \Gamma^\mu{}_{\alpha\beta} + A^{\mu\alpha} \Gamma^\nu{}_{\alpha\beta}; \tag{5.65}$$

$$\nabla_\beta B^\mu{}_\nu = B^\mu{}_{\nu,\beta} + B^\alpha{}_\nu \Gamma^\mu{}_{\alpha\beta} - B^\mu{}_\alpha \Gamma^\alpha{}_{\nu\beta}. \tag{5.66}$$

Inspect these closely: they are *very* systematic. Simply throw in one Γ term for each index; a raised index is treated like a vector and a lowered one like a one-form. The geometrical meaning of Eq. (5.64) is that $\nabla_\beta T_{\mu\nu}$ is a component of the $\binom{0}{3}$ tensor **∇T**, where **T** is a $\binom{0}{2}$ tensor. Similarly, in Eq. (5.65), **A** is a $\binom{2}{0}$ tensor and **∇A** is a $\binom{2}{1}$ tensor with components $\nabla_\beta A^{\mu\nu}$.

5.4 Christoffel symbols and the metric

The formalism developed above has not used any properties of the metric tensor to derive covariant derivatives. But the metric must be involved somehow, because it can convert a vector into a one-form, and so it must have something to say about the relationship between their derivatives. In particular, in Cartesian coordinates the components of the one-form and its related vector are *equal*, and since ∇ is just differentiation of components, the components of the covariant derivatives of the one-form and vector must be equal. This means that if $\vec{V}$ is an arbitrary vector and $\tilde{V}$ = **g**($\vec{V}$,) is its related one-form, then in Cartesian coordinates

$$\nabla_\beta \tilde{V} = \mathbf{g}(\nabla_\beta \vec{V}, \quad). \tag{5.67}$$

But Eq. (5.67) is a tensor equation, so it must be valid in *all* coordinates. We conclude that

$$V_{\alpha;\beta} = g_{\alpha\mu} V^\mu{}_{;\beta}, \tag{5.68}$$

which is the component representation of Eq. (5.67).

If the above argument in words wasn't satisfactory, let us go through it again in equations. Let unprimed indices α, β, γ, … denote Cartesian coordinates and primed indices α', β', γ', … denote *arbitrary* coordinates.

We begin with the statement

$$V_{\alpha'} = g_{\alpha'\mu'} V^{\mu'}, \tag{5.69}$$

valid in any coordinate system. But in Cartesian coordinates

$$g_{\alpha\mu} = \delta_{\alpha\mu}, \quad V_\alpha = V^\alpha.$$

Now, also in Cartesian coordinates, the Christoffel symbols vanish, so

$$V_{\alpha;\beta} = V_{\alpha,\beta} \quad \text{and} \quad V^\alpha{}_{;\beta} = V^\alpha{}_{,\beta}.$$

Therefore we conclude

$$V_{\alpha;\beta} = V^\alpha{}_{;\beta}$$

in Cartesian coordinates only. To convert this into an equation valid in all coordinate systems, we note that in Cartesian coordinates

$$V^\alpha{}_{;\beta} = g_{\alpha\mu} V^\mu{}_{;\beta},$$

so that again in Cartesian coordinates we have

$$V_{\alpha;\beta} = g_{\alpha\mu} V^{\mu}{}_{;\beta}.$$

But now this equation *is* a tensor equation, so its validity in one coordinate system implies its validity in all. This is just Eq. (5.68) again:

$$V_{\alpha';\beta'} = g_{\alpha'\mu'} V^{\mu'}{}_{;\beta'} \tag{5.70}$$

This result has far-reaching implications. If we take the β' covariant derivative of Eq. (5.69), we find $V_{\alpha';\beta'} = g_{\alpha'\mu';\beta'} V^{\mu'} + g_{\alpha'\mu'} V^{\mu'}{}_{;\beta'}.$ Comparison of this with Eq. (5.70) shows (since $\vec{V}$ is an arbitrary vector) that we must have

$$g_{\alpha'\mu';\beta'} \equiv 0 \tag{5.71}$$

in all coordinate systems. This is a consequence of Eq. (5.67). In Cartesian coordinates $g_{\alpha\mu;\beta} \equiv g_{\alpha\mu,\beta} = \delta_{\alpha\mu,\beta} \equiv 0$ is a trivial identity. However, in other coordinates it is not obvious, so we shall work it out as a check on the consistency of our formalism.

Using Eq. (5.64) gives (now unprimed indices are general)

$$g_{\alpha\beta;\mu} = g_{\alpha\beta,\mu} - \Gamma^{\nu}{}_{\alpha\mu} g_{\nu\beta} - \Gamma^{\nu}{}_{\beta\mu} g_{\alpha\nu}. \tag{5.72}$$

In polar coordinates let us work out a few examples. Let $\alpha = r$, $\beta = r$, $\mu = r$: $g_{rr;r} = g_{rr,r} - \Gamma^{\nu}{}_{rr} g_{\nu r} - \Gamma^{\nu}{}_{rr} g_{r\nu}.$ Since $g_{rr,r} = 0$ and $\Gamma^{\nu}{}_{rr} = 0$ for all ν, this is trivially zero. Not so trivial is $\alpha = \theta$, $\beta = \theta$, $\mu = r$: $g_{\theta\theta;r} = g_{\theta\theta,r} - \Gamma^{\nu}{}_{\theta r} g_{\nu\theta} - \Gamma^{\nu}{}_{\theta r} g_{\theta\nu}.$ With $g_{\theta\theta} = r^2$, $\Gamma^{\theta}{}_{\theta r} = 1/r$ and $\Gamma^{r}{}_{\theta r} = 0$, this becomes $g_{\theta\theta\,;r} = (r^2)_{,r} - \frac{1}{r}(r^2) - \frac{1}{r}(r^2) = 0.$ So it works, almost magically. But it is important to realize that it is not magic: it follows directly from the facts that $g_{\alpha\beta,\mu} = 0$ in Cartesian coordinates and that $g_{\alpha\beta;\mu}$ are the components of the *same* tensor $\nabla \mathbf{g}$ in arbitrary coordinates.

Perhaps it is useful to pause here to get some perspective on what we have just done. We introduced covariant differentiation in arbitrary coordinates by using our understanding of parallelism in Euclidean space. We then showed that the metric of Euclidean space is covariantly constant: Eq. (5.71). When we go on to curved (Riemannian) spaces we will have to discuss parallelism much more carefully, but Eq. (5.71) will *still* be true, and therefore so will all its consequences, such as those we now go on to describe.

Calculating the Christoffel symbols from the metric

The vanishing of Eq. (5.72) leads to an extremely important result. We see that Eq. (5.72) can be used to determine $g_{\alpha\beta,\mu}$ in terms of $\Gamma^{\mu}{}_{\alpha\beta}$. It turns out that the reverse is also true, that $\Gamma^{\mu}{}_{\alpha\beta}$ can be expressed in terms of $g_{\alpha\beta,\mu}$. This gives an easy way to derive the Christoffel symbols. To show this we first prove a result of some importance in its own right: *in any coordinate system* $\Gamma^{\mu}{}_{\alpha\beta} \equiv \Gamma^{\mu}{}_{\beta\alpha}$. To prove this symmetry consider an arbitrary scalar field ϕ. Its first derivative $\nabla\phi$ is a one-form with components $\phi_{,\beta}$. Its second covariant derivative $\nabla\nabla\phi$ has components $\phi_{,\beta;\alpha}$ and is a $\binom{0}{2}$ tensor. In Cartesian coordinates these components are $\phi_{,\beta,\alpha} := \frac{\partial}{\partial x^{\alpha}} \frac{\partial}{\partial x^{\beta}} \phi$
and we see that they are symmetric in α and β, since partial derivatives commute. But if a tensor is symmetric in one basis it is symmetric in all bases. Therefore

$$\phi_{,\beta;\alpha} = \phi_{,\alpha;\beta} \tag{5.73}$$

in *any* basis. Using the definition, Eq. (5.63) gives

$$\phi_{,\beta,\alpha} - \phi_{,\mu}\Gamma^{\mu}{}_{\beta\alpha} = \phi_{,\alpha,\beta} - \phi_{,\mu}\Gamma^{\mu}{}_{\alpha\beta}$$

in any coordinate system. But again we have

$$\phi_{,\alpha,\beta} = \phi_{,\beta,\alpha}$$

in *any* coordinates, which leaves

$$\Gamma^{\mu}{}_{\alpha\beta}\phi_{,\mu} = \Gamma^{\mu}{}_{\beta\alpha}\phi_{,\mu}$$

for arbitrary ϕ. This proves the assertion

$$\Gamma^{\mu}{}_{\alpha\beta} = \Gamma^{\mu}{}_{\beta\alpha} \quad \text{in any coordinate system.} \tag{5.74}$$

We use this to invert Eq. (5.72) by some advanced index gymnastics. We write three versions of Eq. (5.72) with different permutations of indices:

$$g_{\alpha\beta,\mu} = \Gamma^{\nu}{}_{\alpha\mu} g_{\nu\beta} + \Gamma^{\nu}{}_{\beta\mu} g_{\alpha\nu},$$
$$g_{\alpha\mu,\beta} = \Gamma^{\nu}{}_{\alpha\beta} g_{\nu\mu} + \Gamma^{\nu}{}_{\mu\beta} g_{\alpha\nu},$$
$$-g_{\beta\mu,\alpha} = -\Gamma^{\nu}{}_{\beta\alpha} g_{\nu\mu} - \Gamma^{\nu}{}_{\mu\alpha} g_{\beta\nu}.$$

We add these up and group terms, using the symmetry of **g**, $g_{\beta\nu} = g_{\nu\beta}$:

$$g_{\alpha\beta,\mu} + g_{\alpha\mu,\beta} - g_{\beta\mu,\alpha}$$
$$= (\Gamma^{\nu}{}_{\alpha\mu} - \Gamma^{\nu}{}_{\mu\alpha}) g_{\nu\beta} + (\Gamma^{\nu}{}_{\alpha\beta} - \Gamma^{\nu}{}_{\beta\alpha}) g_{\nu\mu} + (\Gamma^{\nu}{}_{\beta\mu} + \Gamma^{\nu}{}_{\mu\beta}) g_{\alpha\nu}.$$

In this equation the first two terms on the right vanish by the symmetry of Γ, Eq. (5.74), and we get

$$g_{\alpha\beta,\mu} + g_{\alpha\mu,\beta} - g_{\beta\mu,\alpha} = 2 g_{\alpha\nu} \Gamma^{\nu}{}_{\beta\mu}.$$

We are almost there. Dividing by 2, multiplying by $g^{\alpha\gamma}$ (with summation implied on α) and using $g^{\alpha\gamma} g_{\alpha\nu} \equiv \delta^{\gamma}{}_{\nu}$
gives

$$\frac{1}{2} g^{\alpha\gamma} (g_{\alpha\beta,\mu} + g_{\alpha\mu,\beta} - g_{\beta\mu,\alpha}) = \Gamma^{\gamma}{}_{\beta\mu}. \tag{5.75}$$

This is the expression of the Christoffel symbols in terms of the partial derivatives of the components of **g**. In polar coordinates, for example,

$$\Gamma^{\theta}{}_{r\theta} = \tfrac{1}{2} g^{\alpha\theta} (g_{\alpha r,\theta} + g_{\alpha\theta,r} - g_{r\theta,\alpha}).$$

Since $g^{r\theta} = 0$ and $g^{\theta\theta} = r^{-2}$, we have

$$\Gamma^{\theta}{}_{r\theta} = \frac{1}{2r^2} (g_{\theta r,\theta} + g_{\theta\theta,r} - g_{r\theta,\theta})$$
$$= \frac{1}{2r^2} g_{\theta\theta,r} = \frac{1}{2r^2} (r^2)_{,r} = \frac{1}{r}.$$

This is the same value for $\Gamma^{\theta}{}_{r\theta}$, as we derived earlier. This method of computing $\Gamma^{\alpha}{}_{\beta\mu}$ is so useful that it is well worth committing Eq. (5.75) to memory. It will be exactly the same in curved spaces.

The tensorial nature of $\Gamma^{\alpha}{}_{\beta\mu}$

Since $\vec{e}_\alpha$ is a vector, $\nabla\vec{e}_\alpha$ is a $\binom{1}{1}$ tensor whose components are $\Gamma^{\mu}{}_{\alpha\beta}$. Here α is fixed and μ and β are the component indices: changing α changes the tensor $\nabla\vec{e}_\alpha$, while changing μ or β changes only the component under discussion. So it is possible to regard μ and β as component indices and α as a label giving the particular tensor referred to. There is one such tensor for each basis vector $\vec{e}_\alpha$. However, this is not terribly useful, since under a change of coordinates the basis changes and the important quantities in the new sys-tem are the *new* tensors $\nabla\vec{e}_{\beta'}$ which are obtained from the old ones $\nabla\vec{e}_\alpha$ in a complicated way: they are *different* tensors, not just different components of the same tensor. So the set $\Gamma^\mu{}_{\alpha\beta}$ in one frame is not obtained by a simple tensor transformation from the set $\Gamma^{\mu'}{}_{\alpha'\beta'}$ of another frame. The easiest example of this is Cartesian coordinates, where $\Gamma^\alpha{}_{\beta\mu} \equiv 0$, while they are not zero in other frames. So in many books it is said that $\Gamma^\mu{}_{\alpha\beta}$ are not components of tensors. As we have seen, this is not strictly true: $\Gamma^\mu{}_{\alpha\beta}$ are the (μ, β) components of a *set* of $\binom{1}{1}$ tensors $\nabla\vec{e}_\alpha$. But there is no single $\binom{1}{2}$ tensor whose components are $\Gamma^\mu{}_{\alpha\beta}$, so expressions like $\Gamma^\mu{}_{\alpha\beta}V^\alpha$ are not components of a single tensor, either. The combination $V^\beta{}_{,\alpha} + V^\mu\Gamma^\beta{}_{\mu\alpha}$ *is* a component of a single tensor $\nabla\vec{V}$.

5.5 Noncoordinate bases

In this whole discussion we have generally assumed that the non-Cartesian basis vectors were generated by a coordinate transformation from (x, y) to some (ξ, η). However, as we shall show below, not every field of basis vectors can be obtained in this way, and we shall have to look carefully at our results to see which need modification (few actually do). We will almost never use noncoordinate bases in our work in this course, but they are frequently encountered in the standard references on curved coordinates in flat space, so we should pause to take a brief look at them now.

Polar coordinate basis

The basis vectors for our polar coordinate system were defined by

$$\vec{e}_{\alpha'} = \Lambda^\beta{}_{\alpha'}\vec{e}_\beta,$$

where primed indices refer to polar coordinates and unprimed to Cartesian. Moreover, we had

$$\Lambda^{\beta}{}_{\alpha'} = \partial x^{\beta}/\partial x^{\alpha'},$$

where we regard the Cartesian coordinates $\{x^{\beta}\}$ as functions of the polar coordinates $\{x^{\alpha'}\}$. We found that $\vec{e}_{\alpha'} \cdot \vec{e}_{\beta'} \equiv g_{\alpha'\beta'} \neq \delta_{\alpha'\beta'}$, *i.e.* that these basis vectors are *not* unit vectors.

Polar unit basis

Often it is convenient to work with *unit* vectors. A simple set of unit vectors derived from the polar coordinate basis is:

$$\vec{e}_{\hat{r}} = \vec{e}_r, \vec{e}_{\hat{\theta}} = \frac{1}{r}\vec{e}_{\theta}, \tag{5.76}$$

with a corresponding unit one-form basis

$$\tilde{\omega}^{\hat{r}} = \tilde{\mathrm{d}}r, \quad \tilde{\omega}^{\hat{\theta}} = r\tilde{\mathrm{d}}\theta. \tag{5.77}$$

The student should verify that

$$\left.\begin{aligned} \vec{e}_{\hat{\alpha}} \cdot \vec{e}_{\hat{\beta}} \equiv g_{\hat{\alpha}\hat{\beta}} = \delta_{\hat{\alpha}\hat{\beta}}, \\ \tilde{\omega}^{\hat{\alpha}} \cdot \tilde{\omega}^{\hat{\beta}} \equiv g^{\hat{\alpha}\hat{\beta}} = \delta^{\hat{\alpha}\hat{\beta}} \end{aligned}\right\} \tag{5.78}$$

so these constitute orthonormal bases for the vectors and one-forms. Our notation, which is fairly standard, is to use a ‘caret’ or ‘hat’, ˆ, above an index to denote an orthornormal basis. Now, the question arises, do there exist coordinates (ξ, η) such that

$$\vec{e}_{\hat{r}} = \vec{e}_{\xi} = \frac{\partial x}{\partial \xi}\,\vec{e}_x + \frac{\partial y}{\partial \xi}\,\vec{e}_y \tag{5.79a}$$

and

$$\vec{e}_{\hat{\theta}} = \vec{e}_{\eta} = \frac{\partial x}{\partial \eta}\,\vec{e}_x + \frac{\partial y}{\partial \eta}\,\vec{e}_y? \tag{5.79b}$$

If so, then $\{\vec{e}_{\hat{r}}, \vec{e}_{\hat{\theta}}\}$ are the basis for the coordinates (ξ, η) and so can be called a coordinate basis; if such (ξ, η) can be shown not to exist, then these vectors are a noncoordinate basis. The question is actually more easily answered if we look at the basis one-forms. Thus, we seek (ξ, η) such that

$$
\left.\begin{aligned}
\tilde{\omega}^{\hat{r}} &= \tilde{\mathrm{d}}\xi = \partial\xi/\partial x\,\tilde{\mathrm{d}}x + \partial\xi/\partial y\,\tilde{\mathrm{d}}y,\\
\tilde{\omega}^{\hat{\theta}} &= \tilde{\mathrm{d}}\eta = \partial\eta/\partial x\,\tilde{\mathrm{d}}x + \partial\eta/\partial y\,\tilde{\mathrm{d}}y.
\end{aligned}\right\}
\tag{5.80}
$$

Since we know $\tilde{\omega}^{\hat{r}}$ and $\tilde{\omega}^{\hat{\theta}}$ in terms of $\tilde{\mathrm{d}}r$ and $\tilde{\mathrm{d}}\theta$, we have, from Eqs. (5.26) and (5.27),

$$
\left.\begin{aligned}
\tilde{\omega}^{\hat{r}} &= \tilde{\mathrm{d}}r = \cos\theta\,\tilde{\mathrm{d}}x + \sin\theta\,\tilde{\mathrm{d}}y,\\
\tilde{\omega}^{\hat{\theta}} &= r\,\tilde{\mathrm{d}}\theta = -\sin\theta\,\tilde{\mathrm{d}}x + \cos\theta\,\tilde{\mathrm{d}}y.
\end{aligned}\right\}
\tag{5.81}
$$

(The orthonormality of $\tilde{\omega}^{\hat{r}}$ and $\tilde{\omega}^{\hat{\theta}}$ are obvious here.) Thus if (ξ, η) exist, we have

$$
\frac{\partial\eta}{\partial x} = -\sin\theta, \quad \frac{\partial\eta}{\partial y} = \cos\theta. \tag{5.82}
$$

If this were true, then the mixed derivatives would be equal:

$$
\frac{\partial}{\partial y}\frac{\partial\eta}{\partial x} = \frac{\partial}{\partial x}\frac{\partial\eta}{\partial y}. \tag{5.83}
$$

This would imply

$$
\frac{\partial}{\partial y}(-\sin\theta) = \frac{\partial}{\partial x}(\cos\theta) \tag{5.84}
$$

or

$$
\frac{\partial}{\partial y}\left(\frac{y}{\surd(x^2+y^2)}\right) + \frac{\partial}{\partial x}\left(\frac{x}{\surd(x^2+y^2)}\right) = 0.
$$

This is certainly *not* true. Therefore ξ and η do *not* exist: we have a noncoordinate basis. (If this manner of proof is surprising, try it on $\tilde{\mathrm{d}}r$ and $\tilde{\mathrm{d}}\theta$ themselves.) In textbooks that deal with vector calculus in curvilinear coordinates, almost all use the unit orthonormal basis rather than the coordinate basis. Thus, for polar coordinates, if a vector has components in the *coordinate* basis *PC*,

$$
\vec{V} \underset{PC}{\longrightarrow} (a, b) = \{V^{\alpha'}\}, \tag{5.85}
$$

then it has components in the *orthonormal* basis *PO*

$$
\vec{V} \underset{PO}{\longrightarrow} (a, rb) = \{V^{\hat{\alpha}}\}. \tag{5.86}
$$

So if, for example, the books calculate the divergence of the vector, they obtain, instead of our Eq. (5.56),

$$\nabla \cdot V = \frac{1}{r}\frac{\partial}{\partial r}(r\, V^{\hat{r}}) + \frac{1}{r}\frac{\partial}{\partial \theta}V^{\hat{\theta}}. \tag{5.87}$$

The difference between Eqs. (5.56) and (5.87) is purely a matter of the basis for $\vec{V}$.

General remarks on noncoordinate bases

The principal differences between coordinate and noncoordinate bases arise from the following. Consider an arbitrary scalar field ϕ and the number $\tilde{d}\phi(\vec{e}_\mu)$, where $\vec{e}_\mu$ is a basis vector of some arbitrary basis. We have used the notation

$$\tilde{d}\phi(\vec{e}_\mu) = \phi_{,\mu}. \tag{5.88}$$

Now, if $\vec{e}_\mu$ is a member of a coordinate basis, then $\tilde{d}\phi(\vec{e}_\mu) = \partial\phi/\partial x^\mu$ and we have, as defined in an earlier chapter,

$$\phi_{,\mu} = \frac{\partial \phi}{\partial x^\mu} : \text{coordinate basis}. \tag{5.89}$$

But if no coordinates exist for $\{\vec{e}_\mu\}$, then Eq. (5.89) must fail. For example, if we let Eq. (5.88) *define* $\phi_{,\hat{\mu}}$, then we have

$$\phi_{,\hat{\theta}} = \frac{1}{r}\frac{\partial \phi}{\partial \theta}. \tag{5.90}$$

In general, we get

$$\nabla_{\hat{\alpha}}\phi \equiv \phi_{,\hat{\alpha}} = \Lambda^{\beta}{}_{\hat{\alpha}}\nabla_\beta \phi = \Lambda^{\beta}{}_{\hat{\alpha}}\frac{\partial \phi}{\partial x^\beta} \tag{5.91}$$

for any coordinate system $\{x^\beta\}$ and noncoordinate basis $\{\vec{e}_{\hat{\alpha}}\}$. It is thus convenient to continue with the notation, Eq. (5.88), and to make the rule that $\phi_{,\mu} = \partial\phi/\partial x^\mu$ only in a coordinate basis.

The Christoffel symbols may be defined just as before

$$\nabla_{\hat{\beta}}\vec{e}_{\hat{\alpha}} = \Gamma^{\hat{\mu}}{}_{\hat{\alpha}\hat{\beta}}\vec{e}_{\hat{\mu}}, \tag{5.92}$$

but now

$$\nabla_{\hat{\beta}} = \Lambda^{\alpha}{}_{\hat{\beta}}\frac{\partial}{\partial x^\alpha}, \tag{5.93}$$

where $\{x^\alpha\}$ is any coordinate system and $\{\vec{e}_{\hat{\beta}}\}$ any basis (coordinate or not). Now, however, we *cannot* prove that $\Gamma^{\hat{\mu}}{}_{\hat{\alpha}\hat{\beta}} = \Gamma^{\hat{\mu}}{}_{\hat{\beta}\hat{\alpha}}$, since that proof used $\phi_{,\hat{\alpha},\hat{\beta}} = \phi_{,\hat{\beta},\hat{\alpha}}$, which was true in a coordinate basis (partial derivatives commute) but is not true otherwise. Hence, also, Eq. (5.75) for $\Gamma^\mu{}_{\alpha\beta}$ in terms of $g_{\alpha\beta,\gamma}$ applies only in a coordinate basis. More general expressions are worked out in Exer. 20, § 5.8.

What is the general reason for the nonexistence of coordinates for a basis? If $\{\tilde{\omega}^{\bar{\alpha}}\}$ is a coordinate one-form basis, then its relation to another one $\{\tilde{\mathrm{d}}x^\alpha\}$ is

$$\tilde{\omega}^{\bar{\alpha}} = \Lambda^{\bar{\alpha}}{}_\beta\, \tilde{\mathrm{d}}x^\beta = \frac{\partial x^{\bar{\alpha}}}{\partial x^\beta}\, \mathrm{d}x^\beta. \tag{5.94}$$

The key point is that $\Lambda^{\bar{\alpha}}{}_\beta$, which is generally a function of position, must actually be the partial derivative $\partial x^{\bar{\alpha}}/\partial x^\beta$ everywhere. Thus we have

$$\frac{\partial}{\partial x^\gamma}\Lambda^{\bar{\alpha}}{}_\beta = \frac{\partial^2 x^{\bar{\alpha}}}{\partial x^\gamma\,\partial x^\beta} = \frac{\partial^2 x^{\bar{\alpha}}}{\partial x^\beta\,\partial x^\gamma} = \frac{\partial}{\partial x^\beta}\Lambda^{\bar{\alpha}}{}_\gamma. \tag{5.95}$$

These ‘integrability conditions’ must be satisfied by all the elements $\Lambda^{\bar{\alpha}}{}_\beta$ in order for $\tilde{\omega}^{\bar{\alpha}}$ to be a coordinate basis. Clearly, we can always choose a transformation matrix for which this fails, thereby generating a noncoordinate basis.

Noncoordinate bases in this book

We shall not have occasion to use such bases very often. Mainly, it is important to understand that they exist, that not every basis is derivable from a coordinate system. The algebra of coordinate bases is simpler in almost every respect. We may ask why the standard treatments of curvilinear coordinates in vector calculus, then, stick to orthonormal bases. The reason is that in such a basis in Euclidean space, the metric has components $\delta_{\alpha\beta}$, so the form of the dot product and the equality of vector and one-form components carry over directly from Cartesian coordinates (which have the *only* orthonormal coordinate basis!). In order to gain the simplicity of coordinate bases for vector and tensor calculus, we have to spend time learning the difference between vectors and one-forms!

5.6 Looking ahead

The work we have done in this chapter has developed almost all the notation and concepts we will need in our study of curved spaces and spacetimes. It is particularly important that the student understands §§ 5.2–5.4 because the mathematics of curvature will be developed by analogy with the development here. What we have to add to all this is a discussion of parallelism, of how to measure the extent to which the Euclidean parallelism axiom fails. This measure is the famous Riemann tensor.

5.7 Further reading

The Eötvös and Pound–Rebka–Snider experiments, and other experimental fundamentals underpinning GR, are discussed by Dicke (1964), Misner *et al.* (1973), Shapiro (1980), and Will (1993, 2006). See Hoffmann (1983) for a less mathematical discussion of the motivation for introducing curvature. For an up-to-date review of the GPS system's use of relativity, see Ashby (2003).

The mathematics of curvilinear coordinates is developed from a variety of points of view in: Abraham and Marsden (1978), Lovelock and Rund (1990), and Schutz (1980*b*).

5.8 Exercises

Repeat the argument that led to Eq. (5.1) under more realistic assumptions: suppose a fraction ε of the kinetic energy of the mass at the bottom can be converted into a photon and sent back up, the remaining energy staying at ground level in a useful form. Devise a perpetual motion engine if Eq. (5.1) is violated.

Explain why a *uniform* external gravitational field would raise no tides on Earth.

(a) Show that the coordinate transformation $(x, y) \to (\xi, \eta)$ with $\xi = x$ and $\eta = 1$ violates Eq. (5.6).

(b) Are the following coordinate transformations good ones? Compute the Jacobian and list any points at which the transformations fail.

(i) $\xi = (x^2 + y^2)^{1/2}$, $\eta = \arctan(y/x)$; (ii) $\xi = \ln x$, $\eta = y$;

(iii) $\xi = \arctan(y/x)$, $\eta = (x^2 + y^2)^{-1/2}$.

A curve is defined by $\{x = f(\lambda), y = g(\lambda), 0 \leqslant \lambda \leqslant 1\}$. Show that the tangent vector $(dx/d\lambda, dy/d\lambda)$ does actually lie tangent to the curve.

Sketch the following curves. Which have the same paths? Find also their tangent vectors where the parameter equals zero.

(a) $x = \sin\lambda$, $y = \cos\lambda$; (b) $x = \cos(2\pi t^2)$, $y = \sin(2\pi t^2 + \pi)$; (c) $x = s$, $y = s + 4$; (d) $x = s^2$, $y = -(s-2)(s+2)$; (e) $x = \mu$, $y = 1$.

Justify the pictures in Fig. 5.5.

Calculate all elements of the transformation matrices $\Lambda^{\alpha'}{}_{\beta}$ and $\Lambda^{\mu}{}_{\nu'}$ for the transformation from Cartesian (x, y) – the unprimed indices – to polar (r, θ) – the primed indices.

(a) (Uses the result of Exer. 7.) Let $f = x^2 + y^2 + 2xy$, and in Cartesian coordinates $\vec{V} \to (x^2 + 3y, y^2 + 3x)$, $\vec{W} \to (1, 1)$. Compute f as a function of r and θ, and find the components of $\vec{V}$ and $\vec{W}$ on the polar basis, expressing them as functions of r and θ.

(b) Find the components of $\tilde{d}f$ in Cartesian coordinates and obtain them in polars (i) by direct calculation in polars, and (ii) by transforming components from Cartesian.

(c) (i) Use the metric tensor in polar coordinates to find the polar components of the one-forms $\tilde{V}$ and $\tilde{W}$ associated with $\vec{V}$ and $\vec{W}$. (ii) Obtain the polar components of $\tilde{V}$ and $\tilde{W}$ by transformation of their Cartesian components.

Draw a diagram similar to Fig. 5.6 to explain Eq. (5.38).

Prove that $\nabla\vec{V}$, defined in Eq. (5.52), is a $\binom{1}{1}$ tensor.

(Uses the result of Exers. 7 and 8.) For the vector field $\vec{V}$ whose Cartesian components are $(x^2 + 3y, y^2 + 3x)$, compute: (a) $V^{\alpha}{}_{,\beta}$ in Cartesian; (b) the transformation $\Lambda^{\mu'}{}_{\alpha}\Lambda^{\beta}{}_{\nu'} V^{\alpha}{}_{,\beta}$ to polars; (c) the components $V^{\mu'}{}_{;\nu'}$ directly in polars using the Christoffel symbols, Eq. (5.45), in Eq. (5.50); (d) the divergence $V^{\alpha}{}_{,\alpha}$ using your results in (a); (e) the divergence $V^{\mu'}{}_{;\mu'}$ using your results in either (b) or (c); (f) the divergence $V^{\mu'}{}_{;\mu'}$ using Eq. (5.56) directly.

For the one-form field $\tilde{p}$ whose Cartesian components are $(x^2 + 3y, y^2 + 3x)$, compute: (a) $p_{\alpha,\beta}$ in Cartesian; (b) the transformation $\Lambda^{\alpha}{}_{\mu'}\Lambda^{\beta}{}_{\nu'}\, p_{\alpha,\beta}$ to polars; (c) the components $p_{\mu';\nu'}$ directly in polars using the Christoffel symbols, Eq. (5.45), in Eq. (5.63).

For those who have done both Exers. 11 and 12, show in polars that $g_{\mu'\alpha'} V^{\alpha'}{}_{;\nu'} = p_{\mu';\nu'}$.

For the tensor whose polar components are $(A^{rr} = r^2, A^{r\theta} = r\sin\theta, A^{\theta r} = r\cos\theta, A^{\theta\theta} = \tan\theta)$, compute Eq. (5.65) in polars for all possible indices.

For the vector whose polar components are $(V^r = 1, V^\theta = 0)$, compute in polars all components of the second covariant derivative $V^{\alpha}{}_{;\mu;\nu}$. (Hint: to find the second derivative, treat the first derivative $V^{\alpha}{}_{;\mu}$ as any $\binom{1}{1}$ tensor: Eq. (5.66).)

16 Fill in all the missing steps leading from Eq. (5.74) to Eq. (5.75).

Discover how each expression $V^{\beta}{}_{,\alpha}$ and $V^{\mu}\Gamma^{\beta}{}_{\mu\alpha}$ separately transforms under a change of coordinates (for $\Gamma^{\beta}{}_{\mu\alpha}$, begin with Eq. (5.44)). Show that neither is the standard tensor law, but that their *sum* does obey the standard law.

Verify Eq. (5.78).

Verify that the calculation from Eq. (5.81) to Eq. (5.84), when repeated for $\tilde{d}r$ and $\tilde{d}\theta$, shows them to be a coordinate basis.

For a noncoordinate basis $\{\vec{e}_{\mu}\}$, define $\nabla_{\vec{e}_{\mu}}\vec{e}_{\nu} - \nabla_{\vec{e}_{\nu}}\vec{e}_{\mu} := c^{\alpha}{}_{\mu\nu}\vec{e}_{\alpha}$ and use this in place of Eq. (5.74) to generalize Eq. (5.75).

Consider the $x - t$ plane of an inertial observer in SR. A certain uniformly accelerated observer wishes to set up an orthonormal coordinate system. By Exer. 21, § 2.9, his world line is

$$t(\lambda) = a\sinh\lambda, \quad x(\lambda) = a\cosh\lambda, \tag{5.96}$$

where a is a constant and $a\lambda$ is his proper time (clock time on his wrist watch).

(a) Show that the spacelike line described by Eq. (5.96) with a as the variable parameter and λ fixed is orthogonal to his world line where they intersect. Changing λ in Eq. (5.96) then generates a *family* of such lines.

(b) Show that Eq. (5.96) defines a transformation from coordinates (t, x) to coordinates (λ, a), which form an *orthogonal* coordinate system. Draw these coordinates and show that they cover only one half of the original $t - x$ plane. Show that the coordinates are bad on the lines $|x| = |t|$, so they really cover two disjoint quadrants.

(c) Find the metric tensor and all the Christoffel symbols in this coordinate system. This observer will do a perfectly good job, provided that he always uses Christoffel symbols appropriately and sticks to events in his quadrant. In this sense, SR admits accelerated observers. The right-hand quadrant in these coordinates is sometimes called *Rindler space*, and the boundary lines $x = \pm t$ bear some resemblance to the black-hole horizons we will study later.

Show that if $U^{\alpha}\nabla_{\alpha}V^{\beta} = W^{\beta}$, then $U^{\alpha}\nabla_{\alpha}V_{\beta} = W_{\beta}$.

This has been tested experimentally to extremely high precision in the so-called Eötvös experiment. See Dicke (1964).

B. Riemann (1826–66) was the first to publish a detailed study of the consequences of dropping Euclid’s parallelism axiom.

We shall denote Euclidean vectors by arrows, and we shall use Greek letters for indices (numbered 1 and 2) to denote the fact that the sum is over all possible (i.e. both) values.

6
Curved manifolds

6.1 Differentiable manifolds and tensors

The mathematical concept of a curved space begins (but does not end) with the idea of a *manifold*. A manifold is essentially a continuous space which looks locally like Euclidean space. To the concept of a manifold is added the idea of curvature itself. The introduction of curvature into a manifold will be the subject of subsequent sections. First we study the idea of a manifold, which we can regard as just a fancy word for 'space'.

Manifolds

The surface of a sphere is a manifold. So is any m-dimensional 'hyperplane' in an n-dimensional Euclidean space ($m \leqslant n$). More abstractly, the set of all rigid rotations of Cartesian coordinates in three-dimensional Euclidean space will be shown below to be a manifold. Basically, a manifold is any set that can be continuously parametrized. The number of independent parameters is the *dimension* of the manifold, and the parameters themselves are the *coordinates* of the manifold. Consider the examples just mentioned. The surface of a sphere is 'parametrized' by two coordinates θ and ϕ. The m-dimensional 'hyperplane' has m Cartesian coordinates, and the set of all rotations can be parametrized by the three 'Euler angles', which in effect give the direction of the axis of rotation (two parameters for this) and the amount of rotation (one parameter). So the set of rotations is a manifold: each point is a particular rotation, and the coordinates are the three parameters. It is a three-dimensional manifold. Mathematically, the association of points with the values of their parameters can be thought of as a mapping of points of a manifold into points of the Euclidean space of the correct dimension. This is the meaning of the fact that a manifold looks locally like Euclidean space: it is 'smooth' and has a certain number of dimensions. It must be stressed that the large-scale topology of a manifold may be very different from Euclidean space: the surface of a torus is not Euclidean, even topologically. But locally the correspondence is good: a small patch of the surface of a torus can be mapped 1–1 into the plane tangent to it. This is the way to think of a manifold: it is a space with coordinates, that locally looks Euclidean but that globally can warp, bend, and do almost anything (as long as it stays continuous).

Differential structure

We shall really only consider 'differentiable manifolds'. These are spaces that are continuous and differentiable. Roughly, this means that in the neighborhood of each point in the manifold it is possible to define a smooth map to Euclidean space that preserves derivatives of scalar functions at that point. The surface of a sphere is differentiable everywhere. That of a cone is differentiable except at its apex. Nearly all manifolds of use in physics are differentiable almost everywhere. The curved spacetimes of GR certainly are.

The assumption of differentiability immediately means that we can define one-forms and vectors. That is, in a certain coordinate system on the manifold, the members of the set $\{\phi_{,\alpha}\}$ are the components of the one-form $\tilde{\mathrm{d}}\phi$; and any set of the form $\{a\phi_{,\alpha} + b\psi_{,\alpha}\}$, where a and b are functions, is also a one-form field. Similarly, every curve (with parameter, say, λ) has a tangent vector $\vec{V}$ defined as the linear function that takes the one-form $\tilde{\mathrm{d}}\phi$ into the derivative of ϕ along the curve, $\mathrm{d}\phi/\mathrm{d}\lambda$:

$$\langle\tilde{\mathrm{d}}\phi, \vec{V}\rangle = \vec{V}(\tilde{\mathrm{d}}\phi) = \nabla_{\vec{V}}\phi = \mathrm{d}\phi/\mathrm{d}\lambda. \qquad (6.1)$$

Any linear combination of vectors is also a vector. Using the vectors and one-forms so defined, we can build up the whole set of tensors of type $\binom{M}{N}$ just as we did in SR. Since we have not yet picked out any $\binom{0}{2}$ tensor to serve as the metric, there is not yet any correspondence between forms and vectors. Everything else, however, is exactly as we had in SR and in polar coordinates. All of this comes only from differentiability, so the set of all tensors is said to be part of the 'differential structure' of the manifold. We will not have much occasion to use that term.

Review

It is useful here to review the fundamentals of tensor algebra. We can summarize the following rules.

) A tensor *field* defines a tensor at every point.

) Vectors and one-forms are linear operators on each other, producing real numbers. The linearity means:

$$\langle \tilde{p}, a\vec{V} + b\vec{W} \rangle = a\langle \tilde{p}, \vec{V} \rangle + b\langle \tilde{p}, \vec{W} \rangle,$$
$$\langle a\tilde{p} + b\tilde{q}, \vec{V} \rangle = a\langle \tilde{p}, \vec{V} \rangle + b\langle \tilde{q}, \vec{V} \rangle,$$

here *a* and *b* are any scalar fields.

) Tensors are similarly linear operators on one-forms and vectors, producing real numbers.

) If two tensors of the same type have equal components in a given basis, they have equal components in all bases and are said to be identical (or equal, or the same). Only tensors of the same type can be equal. In particular, if a tensor's components are all zero in one basis, they are zero in all, and the tensor is said to be zero.

) A number of manipulations of components of tensor fields are called 'permissible tensor operations' because they produce components of new tensors:

(i) Multiplication by a scalar field produces components of a new tensor of the same type.

(ii) Addition of components of two tensors of the same type gives components of a new tensor of the same type. (In particular, only tensors of the same type can be added.)

(iii) Multiplication of components of two tensors of arbitrary type gives components of a new tensor of the sum of the types, the outer product of the two tensors.

(iv) Covariant differentiation (to be discussed later) of the components of a tensor of type $\binom{N}{M}$ gives components of a tensor of type $\binom{N}{M+1}$.

(v) Contraction on a pair of indices of the components of a tensor of type $\binom{N}{M}$ produces components of a tensor of type $\binom{N-1}{M-1}$. (Contraction is only defined between an upper and lower index.) (6) If an equation is formed using components of tensors combined only by the permissible tensor operations, and if the equation is true in one basis, then it is true in any other. This is a very useful result. It comes from the fact that the equation (from (5) above) is simply an equality between components of two tensors of the same type, which (from (4)) is then true in any system.

6.2 Riemannian manifolds

So far we have not introduced a metric on to the manifold. Indeed, on certain manifolds a metric would be unnecessary or inconvenient for whichever problem is being considered. But in our case the metric is absolutely fundamental, since it will carry the information about the rates at which clocks run and the distances between points, just as it does in SR. A differentiable manifold on which a symmetric $\binom{0}{2}$ tensor field **g** has been singled 2 out to act as the metric at each point is called a Riemannian manifold. (Strictly speaking, only if the metric is positive-definite – that is, **g**($\vec{V}$, $\vec{V}$) > 0 for all $\vec{V}$ ≠ 0 – is it called Riemannian; indefinite metrics, like SR and GR, are called pseudo-Riemannian. This is a distinction that we won't bother to make.) It is important to understand that in picking out a metric we 'add' structure to the manifold; we shall see that the metric completely defines the curvature of the manifold. Thus, by our choosing one metric **g** the manifold gets a certain curvature (perhaps that of a sphere), while a different **g**′ would give it a different curvature (perhaps an ellipsoid of revolution). The differentiable manifold itself is 'primitive': an amorphous collection of points, arranged locally like the points of Euclidean space, but not having any distance relation or shape specified. Giving the metric **g** gives it a specific shape, as we shall see. From now on we shall study Riemannian manifolds, on which a metric **g** is assumed to be defined at every point.

(For completeness we should remark that it is in fact possible to define the notion of curvature on a manifold without introducing a metric (so-called 'affine' manifolds). Some texts actually approach the subject this way. But since the metric is essential in GR, we shall simply study those manifolds whose curvature is defined by a metric.)

The metric and local flatness

The metric, of course, provides a mapping between vectors and one-forms at every point. Thus, given a vector field $\vec{V}(\mathcal{P})$ (which notation means that $\vec{V}$ depends on the position $\mathcal{P}$, where $\mathcal{P}$ is any point), there is a unique one-form field $\tilde{V}(\mathcal{P})$ = **g**($\vec{V}(\mathcal{P})$,). The mapping must be invertible, so that associated with $\tilde{V}$ ($\mathcal{P}$) there is a unique $\vec{V}(\mathcal{P})$. The components of **g** are called $g_{\alpha\beta}$; the components of the inverse matrix are called $g^{\alpha\beta}$. The metric permits raising and lowering of indices in the same way as in SR, which means

$$V_\alpha = g_{\alpha\beta} V^\beta.$$

In general, $\{g_{\alpha\beta}\}$ will be complicated functions of position, so it will not be true that there would be a simple relation between, say, V_0 and V^0 in an arbitrary coordinate system.

Since we wish to study general curved manifolds, we have to allow any coordinate system. In SR we only studied Lorentz (inertial) frames because they were simple. But because gravity prevents such frames from being global, we shall have to allow all coordinates, and hence all coordinate transformations, that are nonsingular. (Nonsingular means, as in § 5.2, that the matrix of the transformation, $\Lambda^{\alpha'}{}_{\beta} \equiv \partial x^{\alpha'}/\partial x^{\beta}$, has an inverse.) Now, the matrix $(g_{\alpha\beta})$ is a symmetric matrix by definition. It is a well-known theorem of matrix algebra (see Exer. 3, § 6.9) that a transformation matrix can always be found that will make any symmetric matrix into a diagonal matrix with each entry on the main diagonal either +1, −1, or zero. The number of +1 entries equals the number of positive eigenvalues of $(g_{\alpha\beta})$, while the number of −1 entries is the number of negative eigenvalues. So if we choose **g** originally to have three positive eigenvalues and one negative, then we can always find a $\Lambda^{\alpha'}{}_{\beta}$ to make the metric components become

$$(g_{\alpha'\beta'}) = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \equiv (\eta_{\alpha\beta}). \tag{6.2}$$

From now on we will use $\eta_{\alpha\beta}$ to denote *only* the matrix in Eq. (6.2), which is of course the metric of SR.

There are two remarks that must be made here. The first is that Eq. (6.2) is only possible if we choose $(g_{\alpha\beta})$ from among the matrices that have three positive and one negative eigenvalues. The sum of the diagonal elements in Eq. (6.2) is called the *signature* of the metric. For SR and GR it is +2. Thus, the fact that we have previously deduced from physical arguments that we can always construct a *local* inertial frame at any event, finds its mathematical representation in Eq. (6.2), that the metric can be transformed into $\eta_{\alpha\beta}$ at that point. This in turn implies that the metric has to have signature +2 if it is to describe a spacetime with gravity.

The second remark is that the matrix $\Lambda^{\alpha'}{}_{\beta}$ that produces Eq. (6.2) at every point may *not* be a coordinate transformation. That is, the set $\{\tilde{\omega}^{\alpha'} = \Lambda^{\alpha'}{}_{\beta}\, \tilde{\mathrm{d}}x^{\beta}\}$ may not be a coordinate basis. By our earlier discussion of noncoordinate bases, it would be a coordinate transformation only if Eq. (5.95) holds:

$$\frac{\partial \Lambda^{\alpha'}{}_{\beta}}{\partial x^{\gamma}} = \frac{\partial \Lambda^{\alpha'}{}_{\gamma}}{\partial x^{\beta}}.$$

In a general gravitational field this will be impossible, because otherwise it would imply the existence of coordinates for which Eq. (6.2) is true everywhere: a global Lorentz frame. However, having found a basis at a particular point $\mathcal{P}$ for which Eq. (6.2) is true, it is possible to find coordinates such that, in the neighborhood of $\mathcal{P}$, Eq. (6.2) is 'nearly' true. This is embodied in the following theorem, whose (rather long) proof is at the end of this section. Choose any point $\mathcal{P}$ of the manifold. A coordinate system $\{x^\alpha\}$ can be found whose origin is at $\mathcal{P}$ and in which:

$$g_{\alpha\beta}(x^{\mu}) = \eta_{\alpha\beta} + 0[(x^{\mu})^2]. \tag{6.3}$$

That is, the metric near $\mathcal{P}$ is approximately that of SR, differences being of second order in the coordinates. From now on we shall refer to such coordinate systems as 'local Lorentz frames' or 'local inertial frames'. Eq. (6.3) can be rephrased in a somewhat more precise way as:

$$g_{\alpha\beta}(\mathcal{P}) = \eta_{\alpha\beta} \quad \text{for all } \alpha, \beta; \tag{6.4}$$

$$\frac{\partial}{\partial x^{\gamma}} g_{\alpha\beta}(\mathcal{P}) = 0 \quad \text{for all } \alpha, \beta, \gamma; \tag{6.5}$$

but generally

$$\frac{\partial^2}{\partial x^{\gamma}\, \partial x^{\mu}} g_{\alpha\beta}(\mathcal{P}) \neq 0$$

for at least some values of α, β, γ , and μ if the manifold is not exactly flat.

The existence of local Lorentz frames is merely the statement that any curved space has a flat space 'tangent' to it at any point. Recall that straight lines in flat spacetime are the world lines of free particles; the absence of first-derivative terms (Eq. (6.5)) in the metric of a curved spacetime will mean that free particles are moving on lines that are locally straight in this coordinate system. This makes such coordinates very useful for us, since the equations of physics will be nearly as simple in them as in flat spacetime, and if constructed by the rules of § 6.1 will be valid in any coordinate system. The proof of this theorem is at the end of this section, and is worth studying.

Lengths and volumes

The metric of course gives a way to define lengths of curves. Let $d\vec{x}$ be a small vector displacement on some curve. Then $d\vec{x}$ has squared length $ds^2 = g_{\alpha\beta}\, dx^\alpha\, dx^\beta$. (Recall that we call this the *line element* of the metric.) If we take the absolute value of this and take its square root, we get a measure of length: $dl \equiv |g_{\alpha\beta}\, dx^\alpha\, dx^\beta|^{1/2}$. Then integrating it gives

$$l = \int_{\substack{\text{along}\\ \text{curve}}} |g_{\alpha\beta}\, dx^\alpha\, dx^\beta|^{1/2} \tag{6.6}$$

$$= \int_{\lambda_0}^{\lambda_1} \left| g_{\alpha\beta} \frac{dx^\alpha}{d\lambda} \frac{dx^\beta}{d\lambda} \right|^{1/2} d\lambda, \tag{6.7}$$

where λ is the parameter of the curve (whose endpoints are λ_0 and λ_1). But since the tangent vector $\vec{V}$ has components $V^\alpha = dx^\alpha/d\lambda$, we finally have:

$$l = \int_{\lambda_0}^{\lambda_1} |\vec{V} \cdot \vec{V}|^{1/2}\, d\lambda \tag{6.8}$$

as the length of the arbitrary curve.

The computation of volumes is very important for integration in spacetime. Here, we mean by 'volume' the four-dimensional volume element we used for integrations in Gauss' law in § 4.4. Let us go to a local Lorentz frame, where we know that a small four-dimensional region has four-volume $dx^0\, dx^1\, dx^2\, dx^3$, where $\{x^\alpha\}$ are the coordinates which at this point give the nearly Lorentz metric, Eq. (6.3). In *any* other coordinate system $\{x^{\alpha'}\}$ it is a well-known result of the calculus of several variables that:

$$dx^0\, dx^1\, dx^2\, dx^3 = \frac{\partial(x^0, x^1, x^2, x^3)}{\partial(x^{0'}, x^{1'}, x^{2'}, x^{3'})} dx^{0'}\, dx^{1'}\, dx^{2'}\, dx^{3'}, \tag{6.9}$$

where the factor $\partial(\ )/\partial(\ )$ is the Jacobian of the transformation from $\{x^{\alpha'}\}$ to $\{x^\alpha\}$, as defined in § 5.2:

$$\begin{aligned}\frac{\partial(x^0, x^1, x^2, x^3)}{\partial(x^{0'}, x^{1'}, x^{2'}, x^{3'})} &= \det \begin{pmatrix} \partial x^0/\partial x^{0'} & \partial x^0/\partial x^{1'} & \cdots \\ \partial x^1/\partial x^{0'} & & \\ \vdots & & \end{pmatrix} \\ &= \det(\Lambda^\alpha{}_{\beta'}).\end{aligned} \tag{6.10}$$

This would be a rather tedious way to calculate the Jacobian, but there is an easier way using the metric. In matrix terminology, the transformation of the metric components is

$$(g) = (\Lambda)(\eta)(\Lambda)^T, \tag{6.11}$$

where (g) is the matrix of $g_{\alpha\beta}$, (η) of $\eta_{\alpha\beta}$, etc., and where 'T' denotes transpose. It follows that the determinants satisfy

$$\det(g) = \det(\Lambda)\det(\eta)\det(\Lambda^T). \tag{6.12}$$

But for any matrix

$$\det(\Lambda) = \det(\Lambda^T), \tag{6.13}$$

and we can easily see from Eq. (6.2) that

$$\det(\eta) = -1. \tag{6.14}$$

Therefore, we get

$$\det(g) = -[\det(\Lambda)]^2. \tag{6.15}$$

Now we introduce the notation

$$g := \det(g_{\alpha'\beta'}), \tag{6.16}$$

which enables us to conclude from Eq. (6.15) that

$$\det(\Lambda^{\alpha}{}_{\beta'}) = (-g)^{1/2}. \tag{6.17}$$

Thus, from Eq. (6.9) we get

$$\begin{aligned} dx^0\, dx^1\, dx^2\, dx^3 &= [-\det(g_{\alpha'\beta'})]^{1/2} dx^{0'}\, dx^{1'}\, dx^{2'}\, dx^{3'} \\ &= (-g)^{1/2} dx^{0'}\, dx^{1'}\, dx^{2'}\, dx^{3'}. \end{aligned} \tag{6.18}$$

This is a very useful result. It is also conceptually an important result because it is the first example of a kind of argument we will frequently employ, an argument that uses locally flat coordinates to generalize our flat-space concepts to analogous ones in curved space. In this case we began with $dx^0\, dx^1\, dx^2\, dx^3 =$

d^4x in a locally flat coordinate system. We argue that this volume element at $\mathcal{P}$ must be the volume physically measured by rods and clocks, since the space is the same as Minkowski space in this small region. We then find that the value of this expression in arbitrary coordinates $\{x^{\alpha'}\}$ is Eq. (6.18), $(-g)^{1/2}\, d^4x'$, which is thus the expression for the true volume in a curved space at any point in any coordinates. We call this the *proper volume element.*

It should not be surprising that the metric comes into it, of course, since the metric measures lengths. We only need remember that in any coordinates the square root of the negative of the determinant of $(g_{\alpha\beta})$ is the thing to multiply by d^4x to get the true, or *proper*, volume element.

Perhaps it would be helpful to quote an example from three dimensions. Here proper volume is $(g)^{1/2}$, since the metric is positive-definite (Eq. (6.14) would have a + sign). In spherical coordinates the line element is $dl^2 = dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta\, d\phi^2$, so the metric is

$$(g_{ij}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & r^2\sin^2\theta \end{pmatrix}. \tag{6.19}$$

Its determinant is $r^4 \sin^2\theta$, so $(g)^{1/2}\, d^3x'$ is

$$r^2 \sin\theta\, dr\, d\theta\, d\phi, \tag{6.20}$$

which we know is the correct volume element in these coordinates.

Proof of the local-flatness theorem

Let $\{x^\alpha\}$ be an arbitrary given coordinate system and $\{x^{\alpha'}\}$ the one which is desired: it reduces to the inertial system at a certain fixed point $\mathcal{P}$. (A point in this four-dimensional manifold is, of course, an event.) Then there is some relation

$$x^\alpha = x^\alpha(x^{\mu'}), \tag{6.21}$$

$$\Lambda^\alpha{}_{\mu'} = \partial x^\alpha / \partial x^{\mu'}. \tag{6.22}$$

Expanding $\Lambda^\alpha{}_{\mu'}$ in a Taylor series about $\mathcal{P}$ (whose coordinates are $x_0^{\mu'}$) gives the transformation at an arbitrary point $\vec{x}$ near $\mathcal{P}$:

$$\Lambda^{\alpha}{}_{\mu'}(\vec{x}) = \Lambda^{\alpha}{}_{\mu'}(\mathcal{P}) + (x^{\gamma'} - x_0^{\gamma'})\frac{\partial \Lambda^{\alpha}{}_{\mu'}}{\partial x^{\gamma'}}(\mathcal{P})$$
$$+ \frac{1}{2}(x^{\gamma'} - x_0^{\gamma'})(x^{\lambda'} - x_0^{\lambda'})\frac{\partial^2 \Lambda^{\alpha}{}_{\mu'}}{\partial x^{\lambda'}\,\partial x^{\gamma'}}(\mathcal{P}) + \cdots,$$
$$= \Lambda^{\alpha}{}_{\mu'}|_{\mathcal{P}} + (x^{\gamma'} - x_0^{\gamma'})\left.\frac{\partial^2 x^{\alpha}}{\partial x^{\gamma'}\,\partial x^{\mu'}}\right|_{\mathcal{P}}$$
$$+ \frac{1}{2}(x^{\gamma'} - x_0^{\gamma'})(x^{\lambda'} - x_0^{\lambda'})\left.\frac{\partial^3 x^{\alpha}}{\partial x^{\lambda'}\,\partial x^{\gamma'}\,\partial x^{\mu'}}\right|_{\mathcal{P}} + \cdots. \tag{6.23}$$

Expanding the metric in the same way gives

$$g_{\alpha\beta}(\vec{x}) = g_{\alpha\beta}|_{\mathcal{P}} + (x^{\gamma'} - x_0^{\gamma'})\left.\frac{\partial g_{\alpha\beta}}{\partial x^{\gamma'}}\right|_{\mathcal{P}}$$
$$+ \frac{1}{2}(x^{\gamma'} - x_0^{\gamma'})(x^{\lambda'} - x_0^{\lambda'})\left.\frac{\partial^2 g_{\alpha\beta}}{\partial x^{\lambda'}\,\partial x^{\gamma'}}\right|_{\mathcal{P}} + \cdots. \tag{6.24}$$

We put these into the transformation,

$$g_{u'v'} = \Lambda^{\alpha}{}_{\mu'}\Lambda^{\beta}{}_{\nu'}g_{\alpha\beta}, \tag{6.25}$$

to obtain

$$\begin{aligned} g_{\mu'\nu'}(\vec{x}) = {} & \Lambda^{\alpha}{}_{\mu'}|_{\mathcal{P}}\Lambda^{\beta}{}_{\nu'}|_{\mathcal{P}}g_{\alpha\beta}|_{\mathcal{P}} \\ & + (x^{\gamma'} - x_0^{\gamma'})[\Lambda^{\alpha}{}_{\mu'}|_{\mathcal{P}}\Lambda^{\beta}{}_{\nu'}|_{\mathcal{P}}g_{\alpha\beta,\gamma'}|_{\mathcal{P}} \\ & + \Lambda^{\alpha}{}_{\mu'}|_{\mathcal{P}}g_{\alpha\beta}|_{\mathcal{P}}\partial^2 x^{\beta}/\partial x^{\gamma'}\partial x^{\nu'}|_{\mathcal{P}} \\ & + \Lambda^{\beta}{}_{\nu'}|_{\mathcal{P}}g_{\alpha\beta}|_{\mathcal{P}}\partial^2 x^{\alpha}/\partial x^{\gamma'}\partial x^{\mu'}|_{\mathcal{P}}] \\ & + \frac{1}{2}(x^{\gamma'} - x_0^{\gamma'})(x^{\lambda'} - x_0^{\lambda'})[\cdots]. \end{aligned} \tag{6.26}$$

Now, we do not know the transformation, Eq. (6.21), but we can define it by its Taylor expansion. Let us count the number of free variables we have for this purpose. The matrix $\Lambda^{\alpha}{}_{\mu'}|_{\mathcal{P}}$ has 16 numbers, all of which are freely specifiable. The array $\{\partial^2 x^{\alpha}/\partial x^{\gamma'}\partial x^{\mu'}|_{\mathcal{P}}\}$ has 4 × 10 = 40 free numbers (not 4 × 4 × 4, since it is *symmetric* in γ' and μ'). The array $\{\partial^3 x^{\alpha}/\partial x^{\lambda'}\partial x^{\gamma'}\partial x^{\mu'}|_{\mathcal{P}}\}$ has 4 × 20 = 80 free variables, since symmetry on *all* rearrangements of λ', γ' and μ' gives only 20 independent arrangements (the general expression for three indices is $n(n+1)(n+2)/3!$, where n is the number of values each index can take, four in our case). On the other hand, $g_{\alpha\beta}|_{\mathcal{P}}$, $g_{\alpha\beta,\gamma'}|_{\mathcal{P}}$ and $g_{\alpha\beta,\gamma'\mu'}|_{\mathcal{P}}$ are all given initially. They have, respectively, 10, 10 × 4 = 40, and 10 × 10 = 100 independent numbers for a fully general metric. The first question is, can we satisfy Eq. (6.4),

$$g_{\mu'\nu'}|_{\mathcal{P}} = \eta_{\mu'\nu'}? \tag{6.27}$$

This can be written as

$$\eta_{\mu'\nu'} = \Lambda^{\alpha}{}_{\mu'}|_{\mathcal{P}}\Lambda^{\beta}{}_{\nu'}|_{\mathcal{P}}g_{\alpha\beta}|_{\mathcal{P}}. \tag{6.28}$$

By symmetry, these are ten equations, which for general matrices are independent. To satisfy them we have 16 free values in $\Lambda^{\alpha}{}_{\mu'}|_{\mathcal{P}}$. The equations can indeed, therefore, be satisfied, leaving six elements of $\Lambda^{\alpha}{}_{\mu'}|_{\mathcal{P}}$ unspecified. These six correspond to the six degrees of freedom in the Lorentz transformations that preserve the form of the metric $\eta_{\mu\nu}$. That is, we can boost by a velocity $\boldsymbol{v}$ (three free parameters) or rotate by an angle θ around a direction defined by two other angles. These add up to six degrees of freedom in $\Lambda^{\alpha}{}_{\mu'}|_{\mathcal{P}}$ that leave the local inertial frame inertial.

The next question is, can we choose the 40 free numbers $\partial\Lambda^{\alpha}{}_{\mu'}/\partial x^{\gamma'}|_{\mathcal{P}}$ in Eq. (6.26) in such a way as to satisfy the 40 independent equations, Eq. (6.5),

$$g_{\alpha'\beta',\mu'}|_{\mathcal{P}} = 0? \tag{6.29}$$

Since 40 equals 40, the answer is yes, just barely. Given the matrix $\Lambda^{\alpha}{}_{\mu'}|_{\mathcal{P}}$, there is one and only one way to arrange the coordinates near $\mathcal{P}$ such that $\Lambda^{\alpha}{}_{\mu',\gamma'}|_{\mathcal{P}}$ has the right values to make $g_{\alpha'\beta',\mu'}|_{\mathcal{P}} = 0$. So there is no extra freedom other than that with which to make local Lorentz transformations.

The final question is, can we make this work at higher order? Can we find 80 numbers $\Lambda^{\alpha}{}_{\mu',\gamma'\lambda'}|_{\mathcal{P}}$ which can make the 100 numbers $g_{\alpha'\beta',\mu'\lambda'}|_{\mathcal{P}} = 0$? The answer, since $80 < 100$, is no. There are, in the general metric, 20 'degrees of freedom' among the second derivatives $g_{\alpha'\beta',\mu'\lambda'}|_{\mathcal{P}}$. Since $100 - 80 = 20$, there will be in general 20 components that cannot be made to vanish.

Therefore we see that a general metric is characterized at any point $\mathcal{P}$ not so much by its value at $\mathcal{P}$ (which can always be made to be $\eta_{\alpha\beta}$), nor by its first derivatives there (which can be made zero), but by the 20 second derivatives there which in general cannot be made to vanish. These 20 numbers will be seen to be the independent components of a tensor which represents the curvature; this we shall show later. In a *flat* space, of course, all 20 vanish. In a general space they do not.

6.3 Covariant differentiation

We now look at the subject of differentiation. By definition, the derivative of a vector field involves the difference between vectors at two different points (in the limit as the points come together). In a curved space the notion of the difference between vectors at different points must be handled with care, since in between the points the space is curved and the idea that vectors at the two points might point in the 'same' direction is fuzzy. However, the local flatness of the Riemannian manifold helps us out. We only need to compare vectors in the limit as they get infinitesimally close together, and we know that we can construct a coordinate system at any point which is as close to being flat as we would like in this same limit. So in a small region the manifold looks flat, and it is then natural to say that the derivative of a vector whose components are constant in this coordinate system is zero at that point. In particular, we say that the derivatives of the basis vectors of a locally inertial coordinate system are zero at $\mathcal{P}$.

Let us emphasize that this is a *definition* of the covariant derivative. For us, its justification is in the physics: the local inertial frame is a frame in which everything is locally like SR, and in SR the derivatives of these basis vectors are zero. This definition immediately leads to the fact that in these coordinates at this point, the covariant derivative of a vector has components given by the partial derivatives of the components (that is, the Christoffel symbols vanish):

$$V^{\alpha}{}_{;\beta} = V^{\alpha}{}_{,\beta} \quad \text{at } \mathcal{P} \text{ in this frame.} \tag{6.30}$$

This is of course also true for any other tensor, including the metric:

$$g_{\alpha\beta;\gamma} = g_{\alpha\beta,\gamma} = 0 \quad \text{at } \mathcal{P}.$$

(The second equality is just Eq. (6.5).) Now, the equation $g_{\alpha\beta;\gamma} = 0$ is true in one frame (the locally inertial one), and is a valid tensor equation; therefore it is true in *any* basis:

$$g_{\alpha\beta;\gamma} = 0 \quad \text{in any basis.} \tag{6.31}$$

This is a very important result, and comes directly from our definition of the covariant derivative. Recalling § 5.4, we see that *if* we have $\Gamma^\mu{}_{\alpha\beta} = \Gamma^\mu{}_{\beta\alpha}$, then Eq. (6.31) leads to Eq. (5.75) for *any* metric:

$$\Gamma^\alpha{}_{\mu\nu} = \frac{1}{2} g^{\alpha\beta}(g_{\beta\mu,\nu} + g_{\beta\nu,\mu} - g_{\mu\nu,\beta}). \tag{6.32}$$

It is left to Exer. 5, § 6.9, to demonstrate, by repeating the flat-space argument now in the locally inertial frame, that $\Gamma^\mu{}_{\beta\alpha}$ is indeed symmetric in any coordinate system, so that Eq. (6.32) is correct in any coordinates. We assumed at the start that at $\mathcal{P}$ in a locally inertial frame, $\Gamma^\alpha{}_{\mu\nu} = 0$. But, importantly, the derivatives of $\Gamma^\alpha{}_{\mu\nu}$ at $\mathcal{P}$ in this frame are not all zero generally, since they involve $g_{\alpha\beta,\gamma\mu}$. This means that even though coordinates can be found in which $\Gamma^\alpha{}_{\mu\nu} = 0$ at a point, these symbols do not generally vanish elsewhere. This differs from flat space, where a coordinate system exists in which $\Gamma^\alpha{}_{\mu\nu} = 0$ everywhere. So we can see that at any given point, the difference between a general manifold and a flat one manifests itself in the derivatives of the Christoffel symbols.

Eq. (6.32) means that, given $g_{\alpha\beta}$, we can calculate $\Gamma^\alpha{}_{\mu\nu}$ everywhere. We can therefore calculate all covariant derivatives, given **g**. To review the formulas:

$$V^\alpha{}_{;\beta} = V^\alpha{}_{,\beta} + \Gamma^\alpha{}_{\mu\beta} V^\mu, \tag{6.33}$$

$$\mathcal{P}_{\alpha;\beta} = \mathcal{P}_{\alpha,\beta} - \Gamma^\mu{}_{\alpha\beta} \mathcal{P}_\mu, \tag{6.34}$$

$$T^{\alpha\beta}{}_{;\gamma} = T^{\alpha\beta}{}_{,\gamma} + \Gamma^\alpha{}_{\mu\gamma} T^{\mu\beta} + \Gamma^\beta{}_{\mu\gamma} T^{\alpha\mu}. \tag{6.35}$$

Divergence formula

Quite often we deal with the divergence of vectors. Given an arbitrary vector field V^α, its divergence is defined by Eq. (5.53),

$$V^\alpha{}_{;\alpha} = V^\alpha{}_{,\alpha} + \Gamma^\alpha{}_{\mu\alpha} V^\mu. \tag{6.36}$$

This formula involves a sum in the Christoffel symbol, which, from Eq. (6.32), is

(6.37)

$$\Gamma^{\alpha}{}_{\mu\alpha} = \frac{1}{2}g^{\alpha\beta}(g_{\beta\mu,\alpha} + g_{\beta\alpha,\mu} - g_{\mu\alpha,\beta})$$
$$= \frac{1}{2}g^{\alpha\beta}(g_{\beta\mu,\alpha} - g_{\mu\alpha,\beta}) + \frac{1}{2}g^{\alpha\beta}g_{\alpha\beta,\mu}. \tag{6.37}$$

This has had its terms rearranged to simplify it: notice that the term in parentheses is antisymmetric in α and β, while it is contracted on α and β with $g^{\alpha\beta}$, which is symmetric. The first term therefore vanishes (see Exer. 26(a), § 3.10) and we find

$$\Gamma^{\alpha}{}_{\mu\alpha} = \frac{1}{2}g^{\alpha\beta}g_{\alpha\beta,\mu}. \tag{6.38}$$

Since $(g^{\alpha\beta})$ is the inverse matrix of $(g_{\alpha\beta})$, it can be shown (see Exer. 7, § 6.9) that the derivative of the determinant g of the matrix $(g_{\alpha\beta})$ is

$$g_{,\mu} = gg^{\alpha\beta}g_{\beta\alpha,\mu}. \tag{6.39}$$

Using this in Eq. (6.38), we find

$$\Gamma^{\alpha}{}_{\mu\alpha} = (\sqrt{-g})_{,\mu}/\sqrt{-g}. \tag{6.40}$$

Then we can write the divergence, Eq. (6.36), as

$$V^{\alpha}{}_{;\alpha} = V^{\alpha}{}_{,\alpha} + \frac{1}{\sqrt{-g}}V^{\alpha}(\sqrt{-g})_{,\alpha} \tag{6.41}$$

or

$$V^{\alpha}{}_{;\alpha} = \frac{1}{\sqrt{-g}}(\sqrt{-g}V^{\alpha})_{,\alpha}. \tag{6.42}$$

This is a very much easier formula to use than Eq. (6.36). It is also important for Gauss' law, where we integrate the divergence over a volume (using, of course, the proper volume element):

$$\int V^{\alpha}{}_{;\alpha}\sqrt{-g}\,\mathrm{d}^4x = \int (\sqrt{-g}V^{\alpha})_{,\alpha}\,\mathrm{d}^4x. \tag{6.43}$$

Since the final term involves simple partial derivatives, the mathematics of

Gauss' law applies to it, just as in SR (§ 4.8):

$$\int (\sqrt{-g}V^{\alpha})_{,\alpha}\,\mathrm{d}^4x = \oint V^{\alpha} n_{\alpha} \sqrt{-g}\,\mathrm{d}^3S. \tag{6.44}$$

This means

$$\int V^{\alpha}{}_{;\alpha} \sqrt{-g}\,\mathrm{d}^4x = \oint V^{\alpha} n_{\alpha} \sqrt{-g}\,\mathrm{d}^3S. \tag{6.45}$$

So Gauss' law does apply on a curved manifold, in the form given by Eq. (6.45). We need to integrate the divergence over proper volume and to use the *proper surface element*, $n_{\alpha}\sqrt{-g}\,\mathrm{d}^3S$, in the surface integral.

6.4 Parallel-transport, geodesics, and curvature

Until now, we have used the local-flatness theorem to develop as much mathematics on curved manifolds as possible without considering the curvature explicitly. Indeed, we have yet to give a precise mathematical definition of curvature. It is important to distinguish two different kinds of curvature: intrinsic and extrinsic. Consider, for example, a cylinder. Since a cylinder is round in one direction, we think of it as curved. This is its *extrinsic* curvature: the curvature it has in relation to the flat three-dimensional space it is part of. On the other hand, a cylinder can be made by rolling a flat piece of paper without tearing or crumpling it, so the *intrinsic* geometry is that of the original paper: it is flat. This means that the distance in the surface of the cylinder between any two points is the same as it was in the original paper; parallel lines remain parallel when continued; in fact, *all* of Euclid's axioms hold for the surface of a cylinder. A two-dimensional 'ant' confined to that surface would decide it was flat; only its global topology is funny, in that going in a certain direction in a straight line brings him back to where he started. The *intrinsic* geometry of an *n*-dimensional manifold considers only the relationships between its points on paths that remain in the manifold (for the cylinder, in the two-dimensional surface). The *extrinsic* curvature of the cylinder comes from considering it as a surface in a space of higher dimension, and asking about the curvature of lines that stay in the surface compared with 'straight' lines that go off it. So *extrinsic* curvature relies on the notion of a higher-dimensional space. In this book, when we talk about the curvature of spacetime, we talk about its *intrinsic* curvature, since it is clear that all world lines are confined to remain in spacetime. Whether or not there is a higher-dimensional space in which our four-dimensional space is an open question that is becoming more and more a subject of discussion within the framework of string theory. The only thing of interest in GR is the intrinsic geometry of spacetime.

The cylinder, as we have just seen, is intrinsically flat; a sphere, on the other hand, has an intrinsically curved surface. To see this, consider Fig. 6.1, in which two neighboring lines begin at A and B perpendicular to the equator, and hence are parallel. When continued as locally straight lines they follow the arc of great circles, and the two lines meet at the pole P. Parallel lines, when continued, do not remain parallel, so the space is not flat.

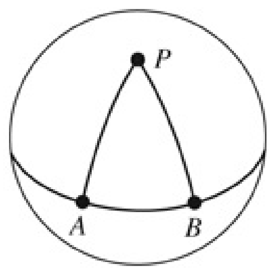


Figure 6.1 A spherical triangle *APB*.

There is an even more striking illustration of the curvature of the sphere. Consider, first, flat space. In Fig. 6.2 a closed path in flat space is drawn, and, starting at A, at each point a vector is drawn parallel to the one at the previous point. This construction is carried around the loop from A to B to C and back to A. The vector finally drawn at A is, of course, parallel to the original one. A completely different thing happens on a sphere! Consider the path shown in Fig. 6.3. Remember, we are drawing the vector as it is seen to a two-dimensional ant on the sphere, so it must always be tangent to the sphere. Aside from that, each vector is drawn as parallel as possible to the previous one. In this loop, A and C are on the equator 90° apart, and B is at the pole. Each arc is the arc of a great circle, and each is 90° long. At A we choose the vector parallel to the equator. As we move up toward B, each new vector is therefore drawn perpendicular to the arc AB. When we get to B, the vectors are tangent to BC. So, going from B to C, we keep drawing tangents to BC. These are perpendicular to the equator at C, and so from C to A the new vectors remain perpendicular to the equator. Thus the vector field has rotated 90° in this construction! Despite the fact that each vector is drawn parallel to its neighbor, the closed loop has caused a discrepancy. Since this doesn't happen in flat space, it must be an effect of the sphere's curvature.

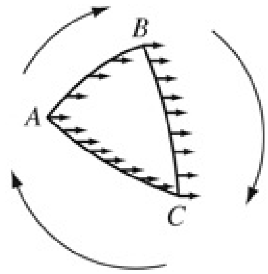


Figure 6.2 A 'triangle' made of curved lines in flat space.

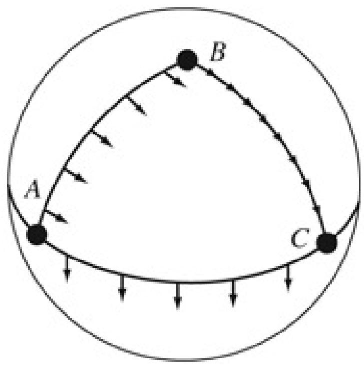


Figure 6.3 Parallel transport around a spherical triangle.

This result has radical implications: on a curved manifold it simply isn't possible to define globally parallel vector fields. We can still define local parallelism, for instance how to move a vector from one point to another, keeping it parallel and of the same length. But the result of such 'parallel transport' from point *A* to point *B* depends on the path taken. We therefore cannot assert that a vector at *A* is or is not parallel to (or the same as) a certain vector at *B*.

Parallel-transport

The construction we have just made on the sphere is called parallel-transport. Suppose a vector field $\vec{V}$ is defined on the sphere, and we examine how it changes along a curve, as in Fig. 6.4. If the vectors $\vec{V}$ at infinitesimally close points of the curve are parallel and of equal length, then $\vec{V}$ is said to be parallel-transported along the curve. It is easy to write down an equation for this. If $\vec{U} = d\vec{x}/d\lambda$ is the tangent to the curve (λ being the parameter along it; $\vec{U}$ is not necessarily normalized), then in a locally inertial coordinate system at a point $\mathcal{P}$

the components of $\vec{V}$ must be constant along the curve at $\mathcal{P}$:

$$\frac{dV^\alpha}{d\lambda} = 0 \quad \text{at}\, \mathcal{P}. \tag{6.46}$$

This can be written as:

$$\frac{dV^\alpha}{d\lambda} = U^\beta V^\alpha{}_{,\beta} = U^\beta V^\alpha{}_{;\beta} = 0 \quad \text{at}\, \mathcal{P}. \tag{6.47}$$

The first equality is the definition of the derivative of a function (in this case V^α) along the curve; the second equality comes from the fact that $\Gamma^\alpha{}_{\mu\nu} = 0$ at $\mathcal{P}$ in these coordinates. But the third equality is a frame-invariant expression and holds in any basis, so it can be taken as a frame-invariant *definition of the parallel-transport of* $\vec{V}$ *along* $\vec{U}$:

$$U^\beta V^\alpha{}_{;\beta} = 0 \Leftrightarrow \frac{d}{d\lambda}\vec{V} = \nabla_{\vec{U}}\vec{V} = 0. \tag{6.48}$$

The last step uses the notation for the derivative along $\vec{U}$ introduced in Eq. (3.67).

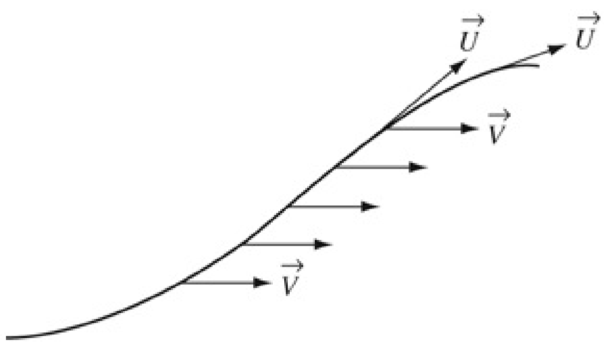


Figure 6.4 Parallel transport of $\vec{V}$ along $\vec{U}$.

Geodesics

The most important curves in flat space are straight lines. One of Euclid’s axioms is that two straight lines that are initially parallel remain parallel when extended. What does he mean by ‘extended’? He *doesn’t* mean ‘continued in such a way that the distance between them remains constant’, because even then they could both bend. What he means is that each line keeps going in the direction it has been going in. More precisely, the tangent to the curve at one point is parallel to the tangent at the previous point. In fact, a straight line in

Euclidean space is the *only* curve that parallel-transports its own tangent vector! In a curved space, we can also draw lines that are 'as nearly straight as possible' by demanding parallel-transport of the tangent vector. These are called *geodesics*:

$$\{\vec{U}\text{is tangent to a geodesic}\} \Leftrightarrow \nabla_{\vec{U}}\vec{U} = 0. \tag{6.49}$$

(Note that in a locally inertial system these lines *are* straight.) In component notation:

$$U^{\beta}U^{\alpha}{}_{;\beta} = U^{\beta}U^{\alpha}{}_{,\beta} + \Gamma^{\alpha}{}_{\mu\beta}U^{\mu}U^{\beta} = 0. \tag{6.50}$$

Now, if we let λ be the parameter of the curve, then $U^{\alpha} = \mathrm{d}x^{\alpha}/\mathrm{d}\lambda$ and $U^{\beta}\partial/\partial x^{\beta} = \mathrm{d}/\mathrm{d}\lambda$:

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{\mathrm{d}x^{\alpha}}{\mathrm{d}\lambda}\right) + \Gamma^{\alpha}{}_{\mu\beta}\frac{\mathrm{d}x^{\mu}}{\mathrm{d}\lambda}\frac{\mathrm{d}x^{\beta}}{\mathrm{d}\lambda} = 0. \tag{6.51}$$

Since the Christoffel symbols $\Gamma^{\alpha}{}_{\mu\beta}$ are known functions of the coordinates $\{x^{\alpha}\}$, this is a nonlinear (quasi-linear), second-order differential equation for $x^{\alpha}(\lambda)$. It has a unique solution when initial conditions at $\lambda = \lambda_0$ are given: $x_0^{\alpha} = x^{\alpha}(\lambda_0)$ and $U_0^{\alpha} = (\mathrm{d}x^{\alpha}/\mathrm{d}\lambda)_{\lambda_0}$. So, by giving an initial position (x_0^{α}) and an initial direction (U_0^{α}), we get a unique geodesic.

Recall that if we change parameter, we change, mathematically speaking, the curve (though not the points it passes through). Now, if λ is a parameter of a geodesic (so that Eq. (6.51) is satisfied), and if we define a new parameter

$$\phi = a\lambda + b, \tag{6.52}$$

where *a* and *b* are *constants* (not depending on position on the curve), then ϕ is also a parameter in which Eq. (6.51) is satisfied: $\dfrac{\mathrm{d}^2x^{\alpha}}{\mathrm{d}\phi^2} + \Gamma^{\alpha}{}_{\mu\beta}\dfrac{\mathrm{d}x^{\mu}}{\mathrm{d}\phi}\dfrac{\mathrm{d}x^{\beta}}{\mathrm{d}\phi} = 0.$ Generally speaking, *only* linear transformations of λ like Eq. (6.52) will give new parameters in which the geodesic equation is satisfied. A parameter like λ and ϕ above is called an *affine* parameter. A curve having the same path as a geodesic but parametrized by a nonaffine parameter is, strictly speaking, not a geodesic curve.

A geodesic is also a curve of *extremal length* between any two points: its length is unchanged to first order in small changes in the curve. The student is urged to prove this by using Eq. (6.7), finding the Euler–Lagrange equations for it to be an extremal for fixed λ_0 and λ_1, and showing that these reduce to Eq. (6.51) when Eq. (6.32) is used. This is a very instructive exercise. We can also show that proper distance along the geodesic is itself an affine parameter (see Exers. 13–15, § 6.9).

6.5 The curvature tensor

At last we are in a position to give a mathematical description of the intrinsic curvature of a manifold. We go back to the curious example of the parallel-transport of a vector around a closed loop, and take it as our *definition* of curvature. Let us imagine in our manifold a very small closed loop (Fig. 6.5) whose four sides are the coordinate lines $x^1 = a$, $x^1 = a + \delta a$, $x^2 = b$, and $x^2 = b + \delta b$. A vector $\vec{V}$ defined at A is parallel-transported to B. The parallel-transport law $\nabla_{\vec{e}_1}\vec{V} = 0$ has the component form

$$\frac{\partial V^\alpha}{\partial x^1} = -\Gamma^\alpha{}_{\mu 1} V^\mu. \tag{6.53}$$

Integrating this from A to B gives

$$\begin{aligned} V^\alpha(B) &= V^\alpha(A) + \int_A^B \frac{\partial V^\alpha}{\partial x^1}\,\mathrm{d}x^1 \\ &= V^\alpha(A) - \int_{x^2=b} \Gamma^\alpha{}_{\mu 1} V^\mu \mathrm{d}x^1, \end{aligned} \tag{6.54}$$

where the notation '$x^2 = b$' under the integral denotes the path AB. Similar transport from B to C to D gives

$$V^\alpha(C) = V^\alpha(B) - \int_{x^1=a+\delta a} \Gamma^\alpha{}_{\mu 2} V^\mu\,\mathrm{d}x^2, \tag{6.55}$$

$$V^\alpha(D) = V^\alpha(C) + \int_{x^2=b+\delta b} \Gamma^\alpha{}_{\mu 1} V^\mu\,\mathrm{d}x^1. \tag{6.56}$$

The integral in the last equation has a different sign because the direction of transport from C to D is in the negative x^1 direction. Similarly, the completion of the loop gives

(6.57)

$$V^\alpha(A_{\text{final}}) = V^\alpha(D) + \int_{x^1=a} \Gamma^\alpha{}_{\mu 2} V^\mu \, dx^2. \tag{6.57}$$

The net change in $V^\alpha(A)$ is a vector δV^α, found by adding Eqs. (6.54)–(6.57):

$$\begin{aligned} \delta V^\alpha &= V^\alpha(A_{\text{final}}) - V^\alpha(A_{\text{initial}}) \\ &= \int_{x^1=a} \Gamma^\alpha{}_{\mu 2} V^\mu dx^2 - \int_{x^1=a+\delta a} \Gamma^\alpha{}_{\mu 2} V^\mu dx^2 \\ &\quad + \int_{x^2=b+\delta b} \Gamma^\alpha{}_{\mu 1} V^\mu dx^1 - \int_{x^2=b} \Gamma^\alpha{}_{\mu 1} V^\mu dx^1. \end{aligned} \tag{6.58}$$

Notice that these would cancel in pairs if $\Gamma^\alpha{}_{\mu\nu}$ and V^μ were constants on the loop, as they would be in flat space. But in curved space they are not, so if we combine the integrals over similar integration variables and work to first order in the separation in the paths, we get to lowest order,

$$\begin{aligned} \delta V^\alpha &\simeq -\int_b^{b+\delta b} \delta a \frac{\partial}{\partial x^1}(\Gamma^\alpha{}_{\mu 2} V^\mu) \, dx^2 \\ &\quad + \int_a^{a+\delta a} \delta b \frac{\partial}{\partial x^2}(\Gamma^\alpha{}_{\mu 1} V^\mu) \, dx^1 \end{aligned} \tag{6.59}$$

$$\approx \delta a \, \delta b \left[-\frac{\partial}{\partial x^1}(\Gamma^\alpha{}_{\mu 2} V^\mu) + \frac{\partial}{\partial x^2}(\Gamma^\alpha{}_{\mu 1} V^\mu) \right]. \tag{6.60}$$

This involves derivatives of Christoffel symbols and of V^α. The derivatives V^α can be eliminated using Eq. (6.53) and its equivalent with 1 replaced by 2. Then Eq. (6.60) becomes

$$\delta V^\alpha = \delta a \, \delta b \left[\Gamma^\alpha{}_{\mu 1,2} - \Gamma^\alpha{}_{\mu 2,1} + \Gamma^\alpha{}_{\nu 2} \Gamma^\nu{}_{\mu 1} - \Gamma^\alpha{}_{\nu 1} \Gamma^\nu{}_{\mu 2} \right] V^\mu. \tag{6.61}$$

(To obtain this, we need to relabel dummy indices in the terms quadratic in Γs.) Notice that this turns out to be just a number times V^μ, summed on μ. Now, the indices 1 and 2 appear because the path was chosen to go along those coordinates. It is antisymmetric in 1 and 2 because the change δV^α would have to have the opposite sign if we went around the loop in the opposite direction (that is, interchanging the roles of 1 and 2). If we used general coordinate lines x^σ and x^λ, we would find

$$
\begin{aligned}
\delta V^\alpha &= \text{change in } V^\alpha \text{ due to transport, first } \delta a\,\vec{e}_\sigma,\text{ then } \delta b\,\vec{e}_\lambda, \\
&\quad \text{then } -\delta a\,\vec{a}_\sigma,\text{ and finally } -\delta b\,\vec{e}_\lambda \\
&= \delta a\,\delta b\left[\Gamma^\alpha{}_{\mu\sigma,\lambda} - \Gamma^\alpha{}_{\mu\lambda,\sigma} + \Gamma^\alpha{}_{\nu\lambda}\Gamma^\nu{}_{\mu\sigma} - \Gamma^\alpha{}_{\nu\sigma}\Gamma^\nu{}_{\mu\lambda}\right]V^\mu.
\end{aligned}
\tag{6.62}
$$

Now, δV^α depends on $\delta a\ \delta b$, the coordinate 'area' of the loop. So it is clear that if the length of the loop in one direction is doubled, δV^α is doubled. This means that δV^α depends *linearly* on $\delta a\vec{e}_\sigma$ and $\delta b\vec{e}_\lambda$. Moreover, it certainly also depends linearly in Eq. (6.62) on V^α itself and on $\tilde{\omega}^\alpha$, which is the basis one-form that gives δV^α from the vector $\delta\,\vec{V}$. Hence we have the following result: if we define

$$
R^\alpha{}_{\beta\mu\nu} := \Gamma^\alpha{}_{\beta\nu,\mu} - \Gamma^\alpha{}_{\beta\mu,\nu} + \Gamma^\alpha{}_{\sigma\mu}\Gamma^\sigma{}_{\beta\nu} - \Gamma^\alpha{}_{\sigma\nu}\Gamma^\sigma{}_{\beta\mu},
\tag{6.63}
$$

then $R^\alpha{}_{\beta\mu\nu}$ must be components of the $\binom{1}{3}$ tensor which, when supplied with arguments $\tilde{\omega}^\alpha$, $\vec{V}$, $\delta a\ \vec{e}_\mu$, $\delta b\ \vec{e}_\nu$, gives δV^α, the component of the change in $\vec{V}$ after parallel-transport around a loop given by $\delta a\ \vec{e}_\mu$ and $\delta b\ \vec{e}_\nu$. This tensor is called the *Riemann curvature tensor* **R**.[1]

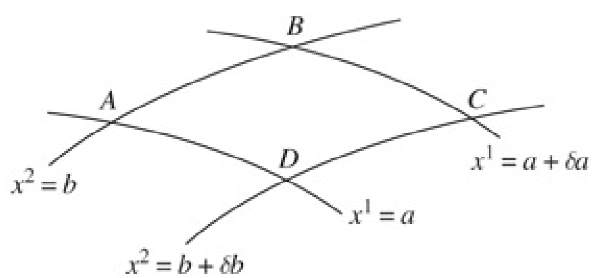


Figure 6.5 Small section of a coordinate grid.

It is useful to look at the components of **R** in a locally inertial frame at a point $\mathcal{P}$. We have $\Gamma^\alpha{}_{\mu\nu} = 0$ at $\mathcal{P}$, but we can find its derivative from Eq. (6.32):

$$
\Gamma^\alpha{}_{\mu\nu,\sigma} = \tfrac{1}{2}g^{\alpha\beta}\left(g_{\beta\mu,\nu\sigma} + g_{\beta\nu,\mu\sigma} - g_{\mu\nu,\beta\sigma}\right).
\tag{6.64}
$$

Since second derivatives of $g_{\alpha\beta}$ don't vanish, we get at $\mathcal{P}$

$$
\begin{aligned}
R^\alpha{}_{\beta\mu\nu} = \tfrac{1}{2}g^{\alpha\sigma}(&g_{\sigma\beta,\nu\mu} + g_{\sigma\nu,\beta\mu} - g_{\beta\nu,\sigma\mu} \\
&- g_{\sigma\beta,\mu\nu} - g_{\sigma\mu,\beta\nu} + g_{\beta\mu,\sigma\nu}).
\end{aligned}
\tag{6.65}
$$

Using the symmetry of $g_{\alpha\beta}$ and the fact that

$$g_{\alpha\beta,\mu\nu} = g_{\alpha\beta,\nu\mu}, \tag{6.66}$$

because partial derivatives always commute, we find at $\mathcal{P}$

$$R^{\alpha}{}_{\beta\mu\nu} = \tfrac{1}{2} g^{\alpha\sigma} \left(g_{\sigma\nu,\beta\mu} - g_{\sigma\mu,\beta\nu} + g_{\beta\mu,\sigma\nu} - g_{\beta\nu,\sigma\mu} \right). \tag{6.67}$$

If we lower the index α, we get (in the locally flat coordinate system at its origin $\mathcal{P}$)

$$R_{\alpha\beta\mu\nu} := g_{\alpha\lambda} R^{\lambda}{}_{\beta\mu\nu} = \tfrac{1}{2} \left(g_{\alpha\nu,\beta\mu} - g_{\alpha\mu,\beta\nu} + g_{\beta\mu,\alpha\nu} - g_{\beta\nu,\alpha\mu} \right). \tag{6.68}$$

In this form it is easy to verify the following identities:

$$R_{\alpha\beta\mu\nu} = -R_{\beta\alpha\mu\nu} = -R_{\alpha\beta\nu\mu} = R_{\mu\nu\alpha\beta}, \tag{6.69}$$

$$R_{\alpha\beta\mu\nu} + R_{\alpha\nu\beta\mu} + R_{\alpha\mu\nu\beta} = 0. \tag{6.70}$$

Thus, $R_{\alpha\beta\mu\nu}$ is antisymmetric on the first pair and on the second pair of indices, and symmetric on exchange of the two pairs. Since Eqs. (6.69) and (6.70) are valid tensor equations true in one coordinate system, they are true in all bases. (Note that an equation like Eq. (6.67) is not a valid tensor equation, since it involves partial derivatives, not covariant ones. Therefore it is true only in the coordinate system in which it was derived.)

It can be shown (Exer. 18, § 6.9) that the various identities, Eqs. (6.69) and (6.70), reduce the number of independent components of $R_{\alpha\beta\mu\nu}$ (and hence of $R^{\alpha}{}_{\beta\mu\nu}$) to 20, in four dimensions. This is, *not* coincidentally, the same number of independent $g_{\alpha\beta,\mu\nu}$ that we found at the end of § 6.2 could not be made to vanish by a coordinate transformation. Thus $R^{\alpha}{}_{\beta\mu\nu}$ characterizes the curvature in a tensorial way.

A *flat* manifold is one which has a global definition of parallelism: a vector can be moved around parallel to itself on an arbitrary curve and will return to its starting point unchanged. This clearly means that

$$R^{\alpha}{}_{\beta\mu\nu} = 0 \Leftrightarrow \text{flat manifold}. \tag{6.71}$$

(Try showing that this is true in polar coordinates for the Euclidean plane.)

An important use of the curvature tensor comes when we examine the consequences of taking two covariant derivatives of a vector field $\vec{V}$. We found in § 6.3 that first derivatives were like flat-space ones, since we could find coordinates in which the metric was flat to first order. But second derivatives are a different story:

$$\begin{aligned} \nabla_\alpha \nabla_\beta V^\mu &= \nabla_\alpha (V^\mu{}_{;\beta}) \\ &= (V^\mu{}_{;\beta})_{,\alpha} + \Gamma^\mu{}_{\sigma\alpha} V^\sigma{}_{;\beta} - \Gamma^\sigma{}_{\beta\alpha} V^\mu{}_{;\sigma}. \end{aligned} \tag{6.72}$$

In locally inertial coordinates whose origin is at $\mathcal{P}$, all the Γs are zero, but their partial derivatives are not. Therefore we have at $\mathcal{P}$

$$\nabla_\alpha \nabla_\beta V^\mu = V^\mu{}_{,\beta\alpha} + \Gamma^\mu{}_{\nu\beta,\alpha} V^\nu. \tag{6.73}$$

Bear in mind that this expression is valid only in this specially chosen coordinates system, and that is true also for Eqs. (6.74) through (6.76) below. These coordinates make the computation easier: consider now Eq. (6.73) with α and β exchanged:

$$\nabla_\beta \nabla_\alpha V^\mu = V^\mu{}_{,\alpha\beta} + \Gamma^\mu{}_{\nu\alpha,\beta} V^\nu. \tag{6.74}$$

If we subtract these, we get the *commutator* of the covariant derivative operators ∇_α and ∇_β, written in the same notation as we would employ in quantum mechanics:

$$\begin{aligned} [\nabla_\alpha, \nabla_\beta] V^\mu &:= \nabla_\alpha \nabla_\beta V^\mu - \nabla_\beta \nabla_\alpha V^\mu \\ &= \left(\Gamma^\mu{}_{\nu\beta,\alpha} - \Gamma^\mu{}_{\nu\alpha,\beta}\right) V^\nu. \end{aligned} \tag{6.75}$$

The terms involving the second derivatives of V^μ drop out here, since

$$V^\mu{}_{,\alpha\beta} = V^\mu{}_{,\beta\alpha}. \tag{6.76}$$

[Let us pause to recall that $V^\mu{}_{,\alpha}$ is the partial derivative of the component V^μ, so by the laws of partial differentiation the partial derivatives must commute. On the other hand, $\nabla_\alpha V^\mu$ is a component of the tensor $\nabla \vec{V}$, and $\nabla_\alpha \nabla_\beta V^\mu$ is a component of $\nabla\nabla\vec{V}$ there is no reason (from differential calculus) why it must be symmetric on α and β. We have proved, by showing that Eq. (6.75) is nonzero, that the double covariant derivative generally is *not* symmetric.] Now, in this frame (where $\Gamma^{\mu}{}_{\alpha\beta} = 0$ at $\mathcal{P}$), we can compare Eq. (6.75) with Eq. (6.63) and see that at $\mathcal{P}$

$$[\nabla_{\alpha}, \nabla_{\beta}]V^{\mu} = R^{\mu}{}_{\nu\alpha\beta}V^{\nu}. \tag{6.77}$$

Now, this is a valid tensor equation, so it is true in *any* coordinate system: the Riemann tensor gives the commutator of covariant derivatives. We can drop the restriction to locally inertial coordinates: they were simply a convenient way of arriving at a general tensor expression for the commutator. What this means is that in curved spaces, we must be careful to know the order in which covariant derivatives are taken: they do not commute. This can be extended to tensors of higher rank. For example, a $\binom{1}{1}$ tensor has

$$[\nabla_{\alpha}, \nabla_{\beta}]F^{\mu}{}_{\nu} = R^{\mu}{}_{\sigma\alpha\beta}F^{\sigma}{}_{\nu} + R_{\nu}{}^{\sigma}{}_{\alpha\beta}F^{\mu}{}_{\sigma}. \tag{6.78}$$

That is, *each* index gets a Riemann tensor on it, and each one comes in with a + sign. (They *must* all have the same sign because raising and lowering indices with **g** is unaffected by ∇_{α}, since $\nabla \mathbf{g} = 0$.) Eq. (6.77) is closely related to our original derivation of the Riemann tensor from parallel-transport around loops, because the parallel-transport problem can be thought of as computing, first the change of $\vec{V}$ in one direction, and then in another, followed by subtracting changes in the reverse order: this is what commuting covariant derivatives also does.

Geodesic deviation

We have often mentioned that in a curved space, parallel lines when extended do not remain parallel. This can now be formulated mathematically in terms of the Riemann tensor. Consider two geodesics (with tangents $\vec{V}$ and $\vec{V}'$) that begin parallel and near each other, as in Fig. 6.6, at points A and A'. Let the affine parameter on the geodesics be called λ.

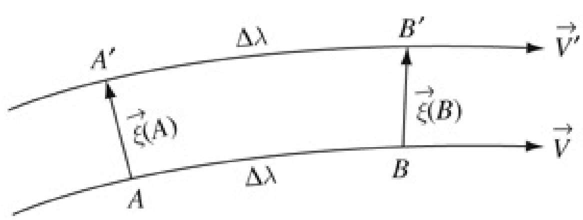


Figure 6.6 A connecting vector $\vec{\xi}$ between two geodesics connects points of the same parameter value.

We define a 'connecting vector' $\vec{\xi}$ which 'reaches' from one geodesic to another, connecting points at equal intervals in λ (i.e. A to A', B to B', etc.). For simplicity, let us adopt a locally inertial coordinate system at A, in which the coordinate x^0 points along the geodesics and advances at the same rate as λ there (this is just a scaling of the coordinate). Then because $V^\alpha = dx^\alpha/d\lambda$, we have at A $V^\alpha = \delta^\alpha_0$. The equation of the geodesic at A is

$$\left.\frac{\mathrm{d}^2x^\alpha}{\mathrm{d}\lambda^2}\right|_A = 0, \tag{6.79}$$

since all Christoffel symbols vanish at A. The Christoffel symbols do not vanish at A', so the equation of the geodesic $\vec{V}'$ at A' is

$$\left.\frac{\mathrm{d}^2x^\alpha}{\mathrm{d}\lambda^2}\right|_{A'} + \Gamma^\alpha{}_{00}(A') = 0, \tag{6.80}$$

where again at A' we have arranged the coordinates so that $V^\alpha = \delta^\alpha_0$. But, since A and A' are separated by $\vec{\xi}$, we have

$$\Gamma^\alpha{}_{00}(A') \cong \Gamma^\alpha{}_{00,\beta}\xi^\beta, \tag{6.81}$$

the right-hand side being evaluated at A. With Eq. (6.80) this gives

$$\left.\frac{\mathrm{d}^2x^\alpha}{\mathrm{d}\lambda^2}\right|_{A'} = -\Gamma^\alpha{}_{00,\beta}\xi^\beta. \tag{6.82}$$

Now, the difference $x^\alpha(\lambda, \text{geodesic } \vec{V}') - x^\alpha(\lambda, \text{geodesic } \vec{V})$ is just the component ξ^α of the vector $\vec{\xi}$. Therefore, at A, we have

$$\frac{\mathrm{d}^2\xi^\alpha}{\mathrm{d}\lambda^2} = \left.\frac{\mathrm{d}^2x^\alpha}{\mathrm{d}\lambda^2}\right|_{A'} - \left.\frac{\mathrm{d}^2x^\alpha}{\mathrm{d}\lambda^2}\right|_A = -\Gamma^\alpha{}_{00,\beta}\xi^\beta. \tag{6.83}$$

This then gives how the components of $\vec{\xi}$ change. But since the coordinates are to some extent arbitrary, we want to have, not merely the second derivative of the component ξ^α, but the full second covariant derivative $\nabla_V\nabla_V\vec{\xi}$. We can use Eq. (6.48) to obtain

$$\begin{aligned}\nabla_V\nabla_V\xi^\alpha &= \nabla_V(\nabla_V\xi^\alpha)\\ &= \frac{\mathrm{d}}{\mathrm{d}\lambda}(\nabla_V\xi^\alpha) = \Gamma^\alpha{}_{\beta 0}(\nabla_V\xi^\beta).\end{aligned} \tag{6.84}$$

Now, using $\Gamma^\alpha{}_{\beta\gamma} = 0$ at A, we have

$$\begin{aligned}\nabla_V \nabla_V \xi^\alpha &= \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{\mathrm{d}}{\mathrm{d}\lambda}\xi^\alpha + \Gamma^\alpha{}_{\beta 0}\xi^\beta\right) + 0 \\ &= \frac{\mathrm{d}^2}{\mathrm{d}\lambda^2}\xi^\alpha + \Gamma^\alpha{}_{\beta 0,0}\xi^\beta\end{aligned} \tag{6.85}$$

at A. (We have also used $\xi^\beta{}_{,0} = 0$ at A, which is the condition that curves begin parallel.) So we get

$$\begin{aligned}\nabla_V \nabla_V \xi^\alpha &= \left(\Gamma^\alpha{}_{\beta 0,0} - \Gamma^\alpha{}_{00,\beta}\right)\xi^\beta \\ &= R^\alpha{}_{00\beta}\xi^\beta = R^\alpha{}_{\mu\nu\beta}V^\mu V^\nu \xi^\beta,\end{aligned} \tag{6.86}$$

where the second equality follows from Eq. (6.63). The final expression is frame invariant, and A was an arbitrary point, so we have, in *any* basis,

$$\nabla_V \nabla_V \xi^\alpha = R^\alpha{}_{\mu\nu\beta}V^\mu V^\nu \xi^\beta. \tag{6.87}$$

Geodesics in flat space maintain their separation; those in curved spaces don't. This is called the equation of geodesic deviation and shows mathematically that the tidal forces of a gravitational field (which cause trajectories of neighboring particles to diverge) can be represented by curvature of a spacetime in which particles follow geodesics.

6.6 Bianchi identities: Ricci and Einstein tensors

Let us return to Eq. (6.63) for the Riemann tensor's components. If we differentiate it with respect to x^λ (just the partial derivative) and evaluate the result in locally inertial coordinates, we find

$$R_{\alpha\beta\mu\nu,\lambda} = \frac{1}{2}\left(g_{\alpha\nu,\beta\mu\lambda} - g_{\alpha\mu,\beta\nu\lambda} + g_{\beta\mu,\alpha\nu\lambda} - g_{\beta\nu,\alpha\mu\lambda}\right). \tag{6.88}$$

From this equation, the symmetry $g_{\alpha\beta} = g_{\beta\alpha}$ and the fact that partial derivatives commute, we can show that

$$R_{\alpha\beta\mu\nu,\lambda} + R_{\alpha\beta\lambda\mu,\nu} + R_{\alpha\beta\nu\lambda,\mu} = 0. \tag{6.89}$$

Since in our coordinates $\Gamma^\mu{}_{\alpha\beta} = 0$ at this point, this equation is *equivalent* to

$$R_{\alpha\beta\mu\nu;\lambda} + R_{\alpha\beta\lambda\mu;\nu} + R_{\alpha\beta\nu\lambda;\mu} = 0. \tag{6.90}$$

But this is a tensor equation, valid in *any* system. It is called the *Bianchi identities*, and will be very important for our work.

The Ricci tensor

Before pursuing the consequences of the Bianchi identities, we shall need to define the Ricci tensor $R_{\alpha\beta}$:

$$R_{\alpha\beta} := R^{\mu}{}_{\alpha\mu\beta} = R_{\beta\alpha}. \tag{6.91}$$

It is the contraction of $R^{\mu}{}_{\alpha\nu\beta}$ on the first and third indices. Other contractions would in principle also be possible: on the first and second, the first and fourth, *etc*. But because $R_{\alpha\beta\mu\nu}$ is antisymmetric on α and β and on μ and ν, all these contractions either vanish identically or reduce to $\pm R_{\alpha\beta}$. Therefore the Ricci tensor is essentially the *only* contraction of the Riemann tensor. Note that Eq. (6.69) implies it is a *symmetric* tensor (Exer. 25, § 6.9).

Similarly, the *Ricci scalar* is defined as

$$R := g^{\mu\nu}R_{\mu\nu} = g^{\mu\nu}g^{\alpha\beta}R_{\alpha\mu\beta\nu}. \tag{6.92}$$

The Einstein tensor

Let us apply the Ricci contraction to the Bianchi identities, Eq. (6.90):

$$g^{\alpha\mu}\left[R_{\alpha\beta\mu\nu;\lambda} + R_{\alpha\beta\lambda\mu;\nu} + R_{\alpha\beta\nu\lambda;\mu}\right] = 0$$

or

$$R_{\beta\nu;\lambda} + (-R_{\beta\lambda;\nu}) + R^{\mu}{}_{\beta\nu\lambda;\mu} = 0. \tag{6.93}$$

To derive this result we need two facts. First, by Eq. (6.31) we have

$$g_{\alpha\beta;\mu} = 0.$$

Since $g^{\alpha\mu}$ is a function only of $g_{\alpha\beta}$ it follows that

$$g^{\alpha\beta}{}_{;\mu} = 0. \tag{6.94}$$

Therefore, $g^{\alpha\mu}$ and $g_{\beta\nu}$ can be taken in and out of covariant derivatives at will:

index-raising and -lowering commutes with covariant differentiation. The second fact is that

$$g^{\alpha\mu}R_{\alpha\beta\lambda\mu;\nu} = -g^{\alpha\mu}R_{\alpha\beta\mu\lambda;\nu} = -R_{\beta\lambda;\nu}, \tag{6.95}$$

accounting for the second term in Eq. (6.93). Eq. (6.93) is called the contracted Bianchi identities. A more useful equation is obtained by contracting again on the indices β and ν:

$$g^{\beta\nu}\left[R_{\beta\nu;\lambda} - R_{\beta\lambda;\nu} + R^{\mu}{}_{\beta\nu\lambda;\mu}\right] = 0$$

or

$$R_{;\lambda} - R^{\mu}{}_{\lambda;\mu} + (-R^{\mu}{}_{\lambda;\mu}) = 0. \tag{6.96}$$

Again the antisymmetry of **R** has been used to get the correct sign in the last term. Note that since R is a scalar, $R_{;\lambda} \equiv R_{,\lambda}$ in all coordinates. Now, Eq. (6.96) can be written in the form

$$(2R^{\mu}{}_{\lambda} - \delta^{\mu}{}_{\lambda}R)_{;\mu} = 0. \tag{6.97}$$

These are the twice-contracted Bianchi identities, often simply also called the Bianchi identities. If we define the symmetric tensor

$$G^{\alpha\beta} \equiv R^{\alpha\beta} - \tfrac{1}{2}g^{\alpha\beta}R = G^{\beta\alpha}, \tag{6.98}$$

then we see that Eq. (6.97) is equivalent to

$$G^{\alpha\beta}{}_{;\beta} = 0. \tag{6.99}$$

The tensor $G^{\alpha\beta}$ is constructed only from the Riemann tensor and the metric, and is automatically divergence free as an identity. It is called the Einstein tensor, since its importance for gravity was first understood by Einstein. [In fact we shall see that the Einstein field equations for GR are

$$G^{\alpha\beta} = 8\pi T^{\alpha\beta}$$

(where $T^{\alpha\beta}$ is the stress-energy tensor). The Bianchi identities then imply

$$T^{\alpha\beta}{}_{;\beta} \equiv 0,$$

which is the equation of local conservation of energy and momentum. But this is looking a bit far ahead.]

6.7 Curvature in perspective

The mathematical machinery for dealing with curvature is formidable. There are many important equations in this chapter, but few of them need to be memorized. It is far more important to understand their derivation and particularly their geometrical interpretation. This interpretation is something we will build up over the next few chapters, but the material already in hand should give the student some idea of what the mathematics means. Let us review the important features of curved spaces.

) We work on Riemannian manifolds, which are smooth spaces with a metric defined on them.

) The metric has signature +2, and there always exists a coordinate system in which, at a single point, we can have

$$\begin{aligned} g_{\alpha\beta} &= \eta_{\alpha\beta}, \\ g_{\alpha\beta,\gamma} &= 0 \Rightarrow \Gamma^{\alpha}_{\beta\gamma} = 0. \end{aligned}$$

) The element of proper volume is

$$|g|^{1/2}\mathrm{d}^4x,$$

where g is the determinant of the matrix of components $g_{\alpha\beta}$.

) The covariant derivative is simply the ordinary derivative in locally inertial coordinates. Because of curvature ($\Gamma^{\alpha}{}_{\beta\gamma,\sigma} \neq 0$) these derivatives do not commute.

) The definition of parallel-transport is that the covariant derivative along the curve is zero. A geodesic parallel-transports its own tangent vector. Its affine parameter can be taken to be the proper distance itself.

) The Riemann tensor is the characterization of the curvature. Only if it vanishes identically is the manifold flat. It has 20 independent components (in four dimensions), and satisfies the Bianchi identities, which are differential equations. The Riemann tensor in a general coordinate system depends on $g_{\alpha\beta}$ and its first and second partial derivatives. The Ricci tensor, Ricci scalar, and Einstein tensor are contractions of the Riemann tensor. In particular, the Einstein tensor is symmetric and of second rank, so it has ten independent components. They satisfy the four differential identities, Eq. (6.99).

6.8 Further reading

The theory of differentiable manifolds is introduced in a large number of books. The following are suitable for exploring the subject further with a view toward its physical applications, particularly outside of relativity: Abraham and Marsden (1978), Bishop and Goldberg (1981), Hermann (1968), Isham (1999), Lovelock and Rund (1990), and Schutz (1980b). Standard mathematical reference works include Kobayashi and Nomizu (1963, 1969), Schouten (1990), and Spivak (1979).

6.9 Exercises

Decide if the following sets are manifolds and say why. If there are exceptional points at which the sets are not manifolds, give them:

(a) phase space of Hamiltonian mechanics, the space of the canonical coordinates and momenta p_i and q^i;

(b) the interior of a circle of unit radius in two-dimensional Euclidean space;

(c) the set of permutations of n objects;

(d) the subset of Euclidean space of two dimensions (coordinates x and y) which is a solution to $xy\,(x^2 + y^2 - 1) = 0$.

Of the manifolds in Exer. 1, on which is it customary to use a metric, and what is that metric? On which would a metric not normally be defined, and why?

It is well known that for any symmetric matrix A (with real entries), there exists a matrix H for which the matrix H^TAH is a diagonal matrix whose entries are the eigenvalues of A.

(a) Show that there is a matrix R such that R^TH^TAHR is the same matrix as H^TAH except with the eigenvalues rearranged in ascending order along the main diagonal from top to bottom.

(b) Show that there exists a third matrix N such that $N^TR^TH^TAHRN$ is a diagonal matrix whose entries on the diagonal are −1, 0, or +1.

(c) Show that if A has an inverse, none of the diagonal elements in (b) is zero.

(d) Show from (a)–(c) that there exists a transformation matrix Λ which produces Eq. (6.2).

Prove the following results used in the proof of the local flatness theorem in § 6.2:

(a) The number of independent values of $\partial^2 x^\alpha/\partial x^{\gamma'}\partial x^{\mu'}|_0$ is 40.
(b) The corresponding number for $\partial^3 x^\alpha/\partial x^{\lambda'}\partial x^{\mu'}\partial x^{\nu'}|_0$ is 80.
(c) The corresponding number for $g_{\alpha\beta,\gamma'\mu'}|_0$ is 100.

(a) Prove that $\Gamma^\mu{}_{\alpha\beta} = \Gamma^\mu{}_{\beta\alpha}$ in any coordinate system in a curved Riemannian space.

(b) Use this to prove that Eq. (6.32) can be derived in the same manner as in flat space.

Prove that the first term in Eq. (6.37) vanishes.

(a) Give the definition of the determinant of a matrix *A* in terms of cofactors of elements.

(b) Differentiate the determinant of an arbitrary 2 × 2 matrix and show that it satisfies Eq. (6.39).

(c) Generalize Eq. (6.39) (by induction or otherwise) to arbitrary *n* × *n* matrices.

Fill in the missing algebra leading to Eqs. (6.40) and (6.42).

Show that Eq. (6.42) leads to Eq. (5.56). Derive the divergence formula for the metric in Eq. (6.19).

A 'straight line' on a sphere is a great circle, and it is well known that the sum of the interior angles of any triangle on a sphere whose sides are arcs of great circles exceeds 180°. Show that the amount by which a vector is rotated by parallel transport around such a triangle (as in Fig. 6.3) equals the excess of the sum of the angles over 180°.

In this exercise we will determine the condition that a vector field $\vec{V}$ can be considered to be globally parallel on a manifold. More precisely, what guarantees that we can find a vector field $\vec{V}$ satisfying the equation

$$(\nabla\vec{V})^\alpha{}_\beta = V^\alpha{}_{;\beta} = V^\alpha{}_{,\beta} + \Gamma^\alpha{}_{\mu\beta}V^\mu = 0?$$

(a) A necessary condition, called the *integrability condition* for this equation, follows from the commuting of partial derivatives. Show that $V^\alpha{}_{,\nu\beta} = V^\alpha{}_{,\beta\nu}$ implies $\left(\Gamma^\alpha{}_{\mu\beta,\nu} - \Gamma^\alpha{}_{\mu\nu,\beta}\right)V^\mu = \left(\Gamma^\alpha{}_{\mu\beta}\Gamma^\mu{}_{\sigma\nu} - \Gamma^\alpha{}_{\mu\nu}\Gamma^\mu{}_{\sigma\beta}\right)V^\sigma$.

(b) By relabeling indices, work this into the form

$$\left(\Gamma^\alpha{}_{\mu\beta,\nu} - \Gamma^\alpha{}_{\mu\nu,\beta} + \Gamma^\alpha{}_{\sigma\nu}\Gamma^\sigma{}_{\mu\beta} - \Gamma^\alpha{}_{\sigma\beta}\Gamma^\sigma{}_{\mu\nu}\right)V^\mu = 0.$$

This turns out to be *sufficient*, as well.

Prove that Eq. (6.52) defines a new affine parameter.

(a) Show that if $\vec{A}$ and $\vec{B}$ are parallel-transported along a curve, then **g**$(\vec{A},\vec{B}) = \vec{A}$

. $\vec{B}$ is constant on the curve.

(b) Conclude from this that if a *geodesic* is spacelike (or timelike or null) somewhere, it is spacelike (or timelike or null) everywhere.

The proper distance along a curve whose tangent is $\vec{V}$ is given by Eq. (6.8). Show that if the curve is a geodesic, then proper length is an affine parameter. (Use the result of Exer. 13.)

Use Exers. 13 and 14 to prove that the proper length of a geodesic between two points is unchanged to first order by small changes in the curve that do not change its endpoints.

(a) Derive Eqs. (6.59) and (6.60) from Eq. (6.58).

(b) Fill in the algebra needed to justify Eq. (6.61).

(a) Prove that Eq. (6.5) implies $g^{\alpha\beta}{}_{,\mu}\,(\mathcal{P}) = 0$.

(b) Use this to establish Eq. (6.64).

(c) Fill in the steps needed to establish Eq. (6.68).

(a) Derive Eqs. (6.69) and (6.70) from Eq. (6.68).

(b) Show that Eq. (6.69) reduces the number of independent components of $R_{\alpha\beta\mu\nu}$ from 4 × 4 × 4 × 4 = 256 to 6 × 7/2 = 21. (Hint: treat *pairs* of indices. Calculate how many independent choices of pairs there are for the first and the second pairs on $R_{\alpha\beta\mu\nu}$.)

(c) Show that Eq. (6.70) imposes only one further relation independent of Eq. (6.69) on the components, reducing the total of independent ones to 20.

Prove that $R^{\alpha}{}_{\beta\mu\nu} = 0$ for polar coordinates in the Euclidean plane. Use Eq. (5.45) or equivalent results.

Fill in the algebra necessary to establish Eq. (6.73).

Consider the sentences following Eq. (6.78). Why does the argument in parentheses *not* apply to the signs in $V^{\alpha}{}_{;\beta} = V^{\alpha}{}_{,\beta} + \Gamma^{\alpha}{}_{\mu\beta}V^{\mu}$ and $V_{\alpha;\beta} = V_{\alpha,\beta} - \Gamma^{\mu}{}_{\alpha\beta}V_{\mu}$?

Fill in the algebra necessary to establish Eqs. (6.84), (6.85), and (6.86).

Prove Eq. (6.88). (Be careful: one cannot simply differentiate Eq. (6.67) since it is valid only at P, not in the neighborhood of P.) **24** Establish Eq. (6.89) from Eq. (6.88).

(a) Prove that the Ricci tensor is the only independent contraction of $R^{\alpha}{}_{\beta\mu\nu}$: all others are multiples of it.

(b) Show that the Ricci tensor is symmetric.

Use Exer. 17(a) to prove Eq. (6.94).

Fill in the algebra necessary to establish Eqs. (6.95), (6.97), and (6.99).

(a) Derive Eq. (6.19) by using the usual coordinate transformation from Cartesian to spherical polars.

(b) Deduce from Eq. (6.19) that the metric of the surface of a sphere of radius r has components ($g_{\theta\theta} = r^2$, $g_{\phi\phi} = r^2 \sin^2 \theta$, $g_{\theta\phi} = 0$) in the usual spherical coordinates.

(c) Find the components $g^{\alpha\beta}$ for the sphere.

In polar coordinates, calculate the Riemann curvature tensor of the sphere of unit radius, whose metric is given in Exer. 28. (Note that in two dimensions there is only *one* independent component, by the same arguments as in Exer. 18(b). So calculate $R_{\theta\phi\theta\phi}$ and obtain all other components in terms of it.) **30** Calculate the Riemann curvature tensor of the cylinder. (Since the cylinder is flat, this should vanish. Use whatever coordinates you like, and make sure you write down the metric properly!)

Show that covariant differentiation obeys the usual product rule, *e.g.* $(V^{\alpha\beta}W_{\beta\gamma})_{;\mu} = V^{\alpha\beta}{}_{;\mu}\, W_{\beta\gamma} + V^{\alpha\beta}W_{\beta\gamma\,;\mu}$. (Hint: use a locally inertial frame.) **32** A four-dimensional manifold has coordinates (u, v, w, p) in which the metric has components $g_{uv} = g_{ww} = g_{pp} = 1$, all other independent components vanishing.

(a) Show that the manifold is flat and the signature is +2.

(b) The result in (a) implies the manifold must be Minkowski spacetime. Find a coordinate transformation to the usual coordinates (t, x, y, z). (You may find it a useful hint to calculate $\vec{e}_v \cdot \vec{e}_v$ and $\vec{e}_u \cdot \vec{e}_u$.) **33** A 'three-sphere' is the three-dimensional surface in four-dimensional Euclidean space (coordinates x, y, z, w), given by the equation $x^2 + y^2 + z^2 + w^2 = r^2$, where r is the radius of the sphere.

(a) Define new coordinates (r, θ, ϕ, χ) by the equations $w = r\cos\chi$, $z = r\sin\chi\cos\theta$, $x = r\sin\chi\sin\theta\cos\phi$, $y = r\sin\chi\sin\theta\sin\phi$. Show that ($\theta$, ϕ, χ) are coordinates for the sphere. These generalize the familiar polar coordinates.

(b) Show that the metric of the three-sphere of radius r has components in these coordinates $g_{\chi\chi} = r^2$, $g_{\theta\theta} = r^2 \sin^2 \chi$, $g_{\phi\phi} = r^2 \sin^2 \chi \sin^2 \theta$, all other components vanishing. (Use the same method as in Exer. 28.) **34** Establish the following identities for a general metric tensor in a general coordinate system. You may find Eqs. (6.39) and (6.40) useful.

(a) $\Gamma^{\mu}{}_{\mu\nu} = \frac{1}{2}(\ln|g|)_{,\nu}$

(b) $g^{\mu\nu}\Gamma^{\alpha}{}_{\mu\nu} = -(g^{\alpha\beta}\sqrt{-g})_{,\beta}/\sqrt{-g}$

(c) for an antisymmetric tensor $F^{\mu\nu}$, $F^{\mu\nu}{}_{;\nu} = (\sqrt{-g}\,F^{\mu\nu})_{,\nu}/\sqrt{-g}$

(d) $g^{\alpha\beta}g_{\beta\mu,\nu} = -g^{\alpha\beta}{}_{,\nu}g_{\beta\mu}$(hint: what is $g^{\alpha\beta}g_{\beta\mu}$?);

(e) $g^{\mu\nu}{}_{,\alpha} = -\Gamma^{\mu}{}_{\beta\alpha}g^{\beta\nu} - \Gamma^{\nu}{}_{\beta\alpha}g^{\mu\beta}$ (hint: use Eq. (6.31)).

Compute 20 independent components of $R_{\alpha\beta\mu\nu}$ for a manifold with line element $ds^2 = -e^{2\Phi}\, dt^2 + e^{2\Lambda}\, dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2)$, where Φ and Λ are arbitrary functions of the coordinate r alone. (First, identify the coordinates and the components $g_{\alpha\beta}$; then compute $g^{\alpha\beta}$ and the Christoffel symbols. Then decide on the indices of the 20 components of $R_{\alpha\beta\mu\nu}$ you wish to calculate, and compute them. Remember that you can deduce the remaining 236 components from those 20.)

A four-dimensional manifold has coordinates (t, x, y, z) and line element
$ds^2 = -(1 + 2\phi)\, dt^2 + (1 - 2\phi)(dx^2 + dy^2 + dz^2)$,
where $|\phi(t, x, y, z)| \ll 1$ everywhere. At any point P with coordinates (t_0, x_0, y_0, z_0), find a coordinate transformation to a locally inertial coordinate system, to first order in ϕ. At what rate does such a frame accelerate with respect to the original coordinates, again to first order in ϕ?

(a) ‘Proper volume’ of a two-dimensional manifold is usually called ‘proper area’. Using the metric in Exer. 28, integrate Eq. (6.18) to find the proper area of a sphere of radius r.

(b) Do the analogous calculation for the three-sphere of Exer. 33.

Integrate Eq. (6.8) to find the length of a circle of constant coordinate θ on a sphere of radius r.

(a) For any two vector fields $\vec{U}$ and $\vec{V}$, their *Lie bracket* is defined to be the vector field $[\vec{U}, \vec{V}]$ with components

$$[\vec{U}, \vec{V}]^{\alpha} = U^{\beta}\nabla_{\beta}V^{\alpha} - V^{\beta}\nabla_{\beta}U^{\alpha}. \tag{6.100}$$

Show that

$$[\vec{U}, \vec{V}] = -[\vec{V}, \vec{U}],$$
$$[\vec{U}, \vec{V}]^{\alpha} = U^{\beta}\,\partial V^{\alpha}/\partial x^{\beta} - V^{\beta}\,\partial U^{\alpha}/\partial x^{\beta}.$$

This is one tensor field in which partial derivatives need not be accompanied by Christoffel symbols!

(b) Show that $[\vec{U}, \vec{V}]$ is a derivative operator on $\vec{V}$ along $\vec{U}$, *i.e.* show that for any scalar f,

$$[\vec{U}, f\vec{V}] = f[\vec{U}, \vec{V}] + \vec{V}(\vec{U}\cdot\nabla f). \tag{6.101}$$

This is sometimes called the *Lie derivative* with respect to $\vec{U}$ and is denoted by

$$[\vec{U}, \vec{V}] := £_{\vec{U}} \vec{V}, \qquad \vec{U} \cdot \nabla f := £_{\vec{U}} f. \tag{6.102}$$

Then Eq. (6.101) would be written in the more conventional form of the Leibnitz rule for the derivative operator $£_{\vec{U}}$:

$$£_{\vec{U}}(f\vec{V}) = f £_{\vec{U}} \vec{V} + \vec{V} £_{\vec{U}} f. \tag{6.103}$$

The result of (a) shows that this derivative operator may be defined without a connection or metric, and is therefore very fundamental. See Schutz (1980*b*) for an introduction.

(c) Calculate the components of the Lie derivative of a one-form field $\tilde{\omega}$ from the knowledge that, for any vector field $\vec{V}, \tilde{\omega}(\vec{V})$ is a scalar like f above, and from the definition that $£_{\vec{U}}\tilde{\omega}$ is a one-form field: $£_{\vec{U}}[\tilde{\omega}(\vec{V})] = (£_{\vec{U}}\tilde{\omega})(\vec{V}) + \tilde{\omega}(£_{\vec{U}}\vec{V})$.

This is the analog of Eq. (6.103).

As with other definitions we have earlier introduced, there is no universal agreement about the overall sign of the Riemann tensor, or even on the placement of its indices. Always check the conventions of whatever book you read.

7

Physics in a curved spacetime

7.1 The transition from differential geometry to gravity

The essence of a physical theory expressed in mathematical form is the identification of the mathematical concepts with certain physically measurable quantities. This must be our first concern when we look at the relation of the concepts of geometry we have developed to the effects of gravity in the physical world. We have already discussed this to some extent. In particular, we have assumed that spacetime is a differentiable manifold, and we have shown that there do not exist global inertial frames in the presence of nonuniform gravitational fields. Behind these statements are the two identifications:

(I) Spacetime (the set of all events) is a four-dimensional manifold with a metric.

(II) The metric is measurable by rods and clocks. The distance along a rod between two nearby points is $|\mathrm{d}\vec{x} \cdot \mathrm{d}\vec{x}|^{1/2}$ and the time measured by a clock that experiences two events closely separated in time is $|-\mathrm{d}\vec{x} \cdot \mathrm{d}\vec{x}|^{1/2}$.

So there do not generally exist coordinates in which $\mathrm{d}\vec{x} \cdot \mathrm{d}\vec{x} = -(\mathrm{d}x^0)^2 + (\mathrm{d}x^1)^2 + (\mathrm{d}x^2)^2 + (\mathrm{d}x^3)^2$ everywhere. On the other hand, we have also argued that such frames *do* exist locally. This clearly suggests a curved manifold, in which coordinates can be found which make the dot product at a particular point look like it does in a Minkowski spacetime.

Therefore we make a further requirement:

(III) The metric of spacetime can be put in the Lorentz form $\eta_{\alpha\beta}$ at any particular event by an appropriate choice of coordinates.

Having chosen this way of representing spacetime, we must do two more things to get a complete theory. First, we must specify how physical objects (particles, electric fields, fluids) behave in a curved spacetime and, second, we need to say how the curvature is generated or determined by the objects in the spacetime.

Let us consider Newtonian gravity as an example of a physical theory. For Newton, spacetime consisted of three-dimensional Euclidean space, repeated endlessly in time. (Mathematically, this is called $R^3 \times R$.) There was no metric on spacetime as a whole manifold, but the Euclidean space had its usual metric and time was measured by a universal clock. Observers with different velocities were all equally valid: this form of relativity was built into Galilean mechanics. Therefore there was no universal standard of rest, and different observers would have different definitions of whether two events occurring at different times happened at the same location. But all observers would agree on simultaneity, on whether two events happened in the same time-slice or not. Thus the 'separation in time' between two events meant the time elapsed between the two Euclidean slices containing the two events. This was independent of the spatial locations of the events, so in Newtonian gravity there was a universal notion of time: all observers, regardless of position or motion, would agree on the elapsed time between two given events. Similarly, the 'separation in space' between two events meant the Euclidean distance between them. If the events were simultaneous, occurring in the same Euclidean time-slice, then this was simple to compute using the metric of that slice, and all observers would agree on it. If the events happened at different times, each observer would take the location of the events in their respective space slices and compute the Euclidean distance between them. The locations would differ for different observers, but again the distance between them would be the same for all observers.

However, in Newtonian theory there was no way to combine the time and distance measures: there was no invariant measure of the length of a general curve that changed position and time as it went along. Without an invariant way of converting times to distances, this was not possible. What Einstein brought to relativity was the invariance of the speed of light, which then permits a unification of time and space measures. Einstein's four-dimensional spacetime has a much simpler structure than Newton's!

Now, within this model of spacetime, Newton gave a law for the behavior of objects that experienced gravitational forces: $\boldsymbol{F} = m\boldsymbol{a}$, where $\boldsymbol{F} = -m\nabla\phi$ for a given gravitational field ϕ. And he also gave a law determining how ϕ is generated: $\nabla^2\Phi = 4\pi G\rho$. These two laws are the ones we must now find analogs for in our relativistic point of view on spacetime. The second one will be dealt with in the next chapter. In this chapter, we ask only how a given metric affects bodies in spacetime.

We have already discussed this for the simple case of particle motion. Since we know that the 'acceleration' of a particle in a gravitational field is independent of its mass, we can go to a freely falling frame in which nearby particles have no acceleration. This is what we have identified as a locally inertial frame. Since freely falling particles have no acceleration in that frame, they follow straight lines, at least locally. But straight lines in a local inertial frame are, of course, the *definition* of geodesics in the full curved manifold. So we have our first postulate on the way particles are affected by the metric:

(IV) *Weak Equivalence Principle*: Freely falling particles move on timelike geodesics of the spacetime.[1]

By 'freely falling' we mean particles unaffected by other forces, such as electric fields, *etc.* All other known forces in physics are distinguished from gravity by the fact that there *are* particles unaffected by them. So the Weak Equivalence Principle (Postulate IV) is a very strong statement, capable of experimental test. And it has been tested, and continues to be tested, to high accuracy. Experiments typically compare the rate of fall of objects that are composed of different materials; current experimental limits bound the fractional differences in acceleration to a few parts in 10^{13} (Will 2006). The WEP is therefore one of the most precisely tested law in all of physics. There are even proposals to test it up to the level of parts in 10^{18} using satellite-borne experiments.

But the WEP refers only to particles. How are, say, fluids affected by a nonflat metric? We need a generalization of (IV):

(IV′) *Einstein Equivalence Principle*: Any local physical experiment not involving gravity will have the same result if performed in a freely falling inertial frame as if it were performed in the flat spacetime of special relativity.

In this case 'local' means that the experiment does not involve fields, such as electric fields, that may extend over large regions and therefore extend outside the domain of validity of the local inertial frame. All of *local* physics is the same in a freely falling inertial frame as it is in special relativity. Gravity introduces nothing new *locally*. All the effects of gravity are felt over extended regions of spacetime. This, too, has been tested rigorously (Will 2006).

This may seem strange to someone used to blaming gravity for making it hard to climb stairs or mountains, or even to get out of bed! But these local effects of gravity are, in Einstein's point of view, really the effects of our being pushed around by the Earth and objects on it. Our 'weight' is caused by the solid Earth exerting forces on us that prevent us from falling freely on a geodesic (weightlessly, through the floor). This is a very reasonable point of view. Consider astronauts orbiting the Earth. At an altitude of some 300 km, they are hardly any further from the center of the Earth than we are, so the strength of the Newtonian gravitational force on them is almost the same as on us. But they are weightless, as long as their orbit prevents them encountering the solid Earth. Once we acknowledge that spacetime has natural curves, the geodesics, and that when we fall on them we are in free fall and feel no gravity, then we can dispose of the Newtonian concept of a gravitational force altogether. We are only following the natural spacetime curve.

The true measure of gravity on the Earth are its tides. These are nonlocal effects, because they arise from the difference of the Moon's Newtonian gravitational acceleration across the Earth, or in other words from the geodesic deviation near the Earth. If the Earth were permanently cloudy, an Earthling would not know about the Moon from its overall gravitational acceleration, since the Earth falls freely: we don't feel the Moon locally. But Earthlings could in principle discover the Moon even without seeing it, by observing and understanding the tides. Tidal forces are the only measurable aspect of gravity.

Mathematically, what the Einstein Equivalence Principle means is, roughly speaking, that if we have a local law of physics that is expressed in tensor notation in SR, then its mathematical form should be the same in a locally inertial frame of a curved spacetime.

This principle is often called the 'comma-goes-to-semicolon rule', because if a law contains derivatives in its special-relativistic form ('commas'), then it has these same derivatives in the local inertial frame. To convert the law into an expression valid in *any* coordinate frame, we simply make the derivatives covariant ('semicolons'). It is an extremely simple way to generalize the physical laws. In particular, it forbids 'curvature coupling': it is conceivable that the correct form of, say, thermodynamics in a curved spacetime would involve the Riemann tensor somehow, which would vanish in SR. Postulate (IV′) would not allow any Riemann-tensor terms in the equations.

As an example of how (IV′) translates into mathematics, we discuss fluid dynamics, which will be our main interest in this course. The law of conservation of particles in SR is expressed as

$$(n\,U^{\alpha})_{,\alpha} = 0, \tag{7.1}$$

where n is the density of particles in the momentarily comoving reference frame (MCRF), and where U^{α} is the four-velocity of a fluid element. In a curved spacetime, at any event, we can find a locally inertial frame comoving momentarily with the fluid element at that event, and define n in exactly the same way. Similarly we can define $\vec{U}$ to be the time basis vector of that frame, just as in SR. Then, according to the Einstein equivalence principle (see Ch. 5), the law of conservation of particles in the locally inertial frame is *exactly* Eq. (7.1). But because the Christoffel symbols are zero at the given event because it is the origin of the locally inertial frame, this is equivalent to

$$(n\,U^{\alpha})_{;\alpha} = 0. \tag{7.2}$$

This form of the law is valid in *all* frames and so allows us to compute the conservation law in any frame and be sure that it is the one implied by the Einstein equivalence principle. We have therefore generalized the law of particle conservation to a curved spacetime. We will follow this method for other laws of physics as we need them.

Is this just a game with tensors, or is there physical content in what we have done? Is it possible that in a curved spacetime the conservation law would actually be something other than Eq. (7.2)? The answer is yes: consider postulating the equation

$$(n\,U^{\alpha})_{;\alpha} = qR^2, \tag{7.3}$$

where R is the Ricci scalar, defined in Eq. (6.92) as the double trace of the Riemann tensor, and where q is a constant. This would also reduce to Eq. (7.1) in SR, since in a flat spacetime the Riemann tensor vanishes. But in curved spacetime, this equation predicts something very different: curvature would either create or destroy particles, according to the sign of the constant q. Thus, both of the previous equations are consistent with the laws of physics in SR. The Einstein equivalence principle asserts that we should generalize Eq. (7.1) in the simplest possible manner, that is to Eq. (7.2). It is of course a matter for experiment, or astronomical observation, to decide whether Eq. (7.2) or Eq. (7.3) is correct. In this book we shall simply make the assumption that is nearly universally made, that the Einstein equivalence principle is correct. There is no observational evidence to the contrary.

Similarly, the law of conservation of entropy in SR is

$$U^{\alpha} S_{,\alpha} = 0. \tag{7.4}$$

Since there are no Christoffel symbols in the covariant derivative of a scalar like S, this law is *unchanged* in a curved spacetime. Finally, conservation of four-momentum is

$$T^{\mu\nu}{}_{,\nu} = 0. \tag{7.5}$$

The generalization is

$$T^{\mu\nu}{}_{;\nu} = 0, \tag{7.6}$$

with the definition

$$T^{\mu\nu} = (\rho + p)U^{\mu} U^{\nu} + p g^{\mu\nu}, \tag{7.7}$$

exactly as before. (Notice that $g^{\mu\nu}$ is the tensor whose components in the local inertial frame equal the flat-space metric tensor $\eta^{\mu\nu}$.)

7.2 Physics in slightly curved spacetimes

To see the implications of (IV′) for the motion of a particle or fluid, we must know the metric on the manifold. Since we have not yet studied the way a metric is generated, we will at this stage have to be content with assuming a form for the metric which we shall derive later. We will see later that for *weak* gravitational fields (where, in Newtonian language, the gravitational potential energy of a particle is much less than its rest-mass energy) the ordinary Newtonian potential ϕ completely determines the metric, which has the form

$$\mathrm{d}s^2 = -(1 + 2\phi)\,\mathrm{d}t^2 + (1 - 2\phi)\,(\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2). \tag{7.8}$$

(The sign of ϕ is chosen negative, so that, far from a source of mass M, we have $\phi = -GM/r$.) Now, the condition above that the field be weak means that $|m\phi| \ll m$, so that $|\phi| \ll 1$. The metric, Eq. (7.8), is really only correct to first order in ϕ, so we shall work to this order from now on.

Let us compute the motion of a freely falling particle. We denote its four-momentum by $\vec{p}$. For all except massless particles, this is $m\vec{U}$, $\vec{U} = \mathrm{d}\vec{x}/\mathrm{d}\tau$. Now, by (IV), the particle's path is a geodesic, and we know that proper time is an affine parameter on such a path. Therefore $\vec{U}$ must satisfy the geodesic equation,

$$\nabla_{\vec{U}} \vec{U} = 0. \tag{7.9}$$

For convenience later, however, we note that any constant times proper time is an affine parameter, in particular τ/m. Then $\mathrm{d}\vec{x}/\mathrm{d}(\tau/m)$ is also a vector satisfying the geodesic equation. This vector is just $m\mathrm{d}\vec{x}/\mathrm{d}\tau = \vec{p}$. So we can also write the equation of motion of the particle as

$$\nabla_{\vec{p}} \vec{p} = 0. \tag{7.10}$$

This equation can also be used for photons, which have a well-defined $\vec{p}$ but no $\vec{U}$ since $m = 0$.

If the particle has a nonrelativistic velocity in the coordinates of Eq. (7.8), we can find an approximate form for Eq. (7.10). First let us consider the zero component of the equation, noting that the ordinary derivative along $\vec{p}$ is m times the ordinary derivative along $\vec{U}$, or in other words $m\ \mathrm{d}/\mathrm{d}\tau$:

$$m\frac{\mathrm{d}}{\mathrm{d}\tau}p^0 + \Gamma^0{}_{\alpha\beta}p^\alpha p^\beta = 0. \tag{7.11}$$

Because the particle has a nonrelativistic velocity we have $p^0 \gg p^1$, so Eq. (7.11) is approximately

$$m\frac{\mathrm{d}}{\mathrm{d}\tau}p^0 + \Gamma^0{}_{00}(p^0)^2 = 0. \tag{7.12}$$

We need to compute $\Gamma^0{}_{00}$:

$$\Gamma^0{}_{00} = \tfrac{1}{2}g^{0\alpha}(g_{\alpha 0,0} + g_{\alpha 0,0} - g_{00,\alpha}). \tag{7.13}$$

Now because $[g_{\alpha\beta}]$ is diagonal, $[g^{\alpha\beta}]$ is also diagonal and its elements are the reciprocals of those of $[g_{\alpha\beta}]$. Therefore $g^{0\alpha}$ is nonzero only when $\alpha = 0$, so Eq. (7.13) becomes

$$\Gamma^0{}_{00} = \tfrac{1}{2}g^{00}g_{00,0} = \frac{1}{2}\frac{1}{-(1+2\phi)}(-2\phi)_{,0}$$
$$= \phi_{,0} + 0(\phi^2). \tag{7.14}$$

To lowest order in the velocity of the particle and in ϕ, we can replace $(p^0)^2$ in the second term of Eq. (7.12) by m^2, obtaining

$$\frac{\mathrm{d}}{\mathrm{d}\tau}p^0 = -m\frac{\partial\phi}{\partial\tau}. \tag{7.15}$$

Since p^0 is the energy of the particle in this frame, this means the energy is conserved unless the gravitational field depends on time. This result is true also in Newtonian theory. Here, however, we must note that p^0 is the energy of the particle with respect to this frame only.

The spatial components of the geodesic equation give the counterpart of the Newtonian $\boldsymbol{F} = m\boldsymbol{a}$. They are

$$p^\alpha p^i{}_{,\alpha} + \Gamma^i{}_{\alpha\beta}p^\alpha p^\beta = 0, \tag{7.16}$$

or, to lowest order in the velocity,

$$m\frac{\mathrm{d}p^i}{\mathrm{d}\tau} + \Gamma^i{}_{00}(p^0)^2 = 0. \tag{7.17}$$

Again we have neglected p^i compared to p^0 in the Γ summation. Consistent with this we can again put $(p^0)^2 = m^2$ to a first approximation and get

$$\frac{\mathrm{d}p^i}{\mathrm{d}\tau} = -m\Gamma^i{}_{00}. \tag{7.18}$$

We calculate the Christoffel symbol:

$$\Gamma^i{}_{00} = \tfrac{1}{2}g^{i\alpha}(g_{\alpha 0,0} + g_{\alpha 0,0} - g_{00,\alpha}). \tag{7.19}$$

Now, since $[g^{\alpha\beta}]$ is diagonal, we can write

$$g^{i\alpha} = (1-2\phi)^{-1}\delta^{i\alpha} \tag{7.20}$$

and get

$$\Gamma^i{}_{00} = \tfrac{1}{2}(1 - 2\phi)^{-1}\delta^{ij}(2g_{j0,0} - g_{00j}), \tag{7.21}$$

where we have changed α to j because δ^{i0} is zero. Now we notice that $g_{j0} \equiv 0$ and so we get

$$\Gamma^i{}_{00} = -\tfrac{1}{2}g_{00,j}\delta^{ij} + 0(\phi^2) \tag{7.22}$$

$$= -\tfrac{1}{2}(-2\phi)_{,j}\delta^{ij}. \tag{7.23}$$

With this the equation of motion, Eq. (7.17), becomes

$$\mathrm{d}p^i/\mathrm{d}\tau = -m\phi_{,j}\delta^{ij}. \tag{7.24}$$

This is the usual equation in Newtonian theory, since the force of a gravitational field is $-m\nabla\phi$. This demonstrates that general relativity predicts the Keplerian motion of the planets, at least so long as the higher-order effects neglected here are too small to measure. We shall see that this is true for most planets, but not for Mercury.

Both the energy-conservation equation and the equation of motion were derived as approximations based on two things: the metric was nearly the Minkowski metric ($|\phi| \ll 1$), and the particle's velocity was nonrelativistic ($p^0 \gg p^i$). These two limits are just the circumstances under which Newtonian gravity is verified, so it is reassuring – indeed, essential – that we have recovered the Newtonian equations. However, there is no magic here. It almost *had* to work, given that we know that particles fall on straight lines in freely falling frames.

We can do the same sort of calculation to verify that the Newtonian equations hold for other systems in the appropriate limit. For instance, the student has an opportunity to do this for the perfect fluid in Exer. 5, § 7.6. Note that the condition that the fluid be nonrelativistic means not only that its velocity is small but also that the random velocities of its particles be nonrelativistic, which means $p \ll \rho$.

This correspondence of our relativistic point of view with the older, Newtonian theory in the appropriate limit is very important. *Any* new theory must make the same predictions as the old theory in the regime in which the old theory was known to be correct. The equivalence principle plus the form of the metric, Eq. (7.8), does this.

7.3 Curved intuition

Although in the appropriate limit our curved-spacetime picture of gravity predicts the same things as Newtonian theory predicts, it is very different from Newton's theory in concept. We must therefore work gradually toward an understanding of its new point of view.

The first difference is the absence of a preferred frame. In Newtonian physics *and* in SR, inertial frames are preferred. Since 'velocity' cannot be measured locally but 'acceleration' can be, both theories single out special classes of coordinate systems for spacetime in which particles which have no physical acceleration (i.e. $d\vec{U}/d\tau = 0$) also have no coordinate acceleration ($d^2x^i/dt^2 = 0$). In our new picture, there is no coordinate system which is inertial everywhere, *i.e.* in which $d^2x^i/dt^2 = 0$ for every particle for which $d\vec{U}/d\tau = 0$. Therefore we have to allow all coordinates on an equal footing. By using the Christoffel symbols we correct coordinate-dependent quantities like d^2x^i/dt^2 to obtain coordinate-independent quantities like $d\vec{U}/d\tau$. Therefore, we need not, and in fact we *should not*, develop coordinate-dependent ways of thinking.

A second difference concerns energy and momentum. In Newtonian physics, SR, and our geometrical gravity theory, each particle has a definite energy and momentum, whose values depend on the frame they are evaluated in. In the latter two theories, energy and momentum are components of a single four-vector $\vec{p}$. In SR, the total four-momentum of a system is the sum of the four-momenta of all the particles, $\sum_i \vec{p}_{(i)}$. But in a curved spacetime, we *cannot* add up vectors that are defined at different points, because we do not know how: two vectors can only be said to be parallel if they are compared at the same point, and the value of a vector at a point to which it has been parallel-transported depends on the curve along which it was moved. So there is *no* invariant way of adding up all the $\vec{p}$s, and so if a system has definable four-momentum, it is not just the simple thing it was in SR.

It turns out that for any system whose spatial extent is bounded (i.e. an isolated system), a total energy and momentum *can* be defined, in a manner which we will discuss later. One way to see that the *total* mass energy of a system should not be the sum of the energies of the particles is that this neglects what in Newtonian language is called its gravitational self-energy, a negative quantity which is the work we gain by assembling the system from isolated particles at infinity. This energy, if it is to be included, cannot be assigned to any particular particle but resides in the geometry itself. The notion of gravitational potential energy, however, is itself not well defined in the new picture: it must in some sense represent the difference between the sum of the energies of the particles and the total mass of the system, but since the sum of the energies of the particles is not well defined, neither is the gravitational potential energy. Only the *total* energy–momentum of a system is, in general, definable, in addition to the four-momentum of individual particles.

7.4 Conserved quantities

The previous discussion of energy may make us wonder what we can say about conserved quantities associated with a particle or system. For a particle, we must realize that gravity, in the old viewpoint, is a 'force', so that a particle's kinetic energy and momentum need not be conserved under its action. In our new viewpoint, then, we cannot expect to find a coordinate system in which the components of $\vec{p}$ are constants along the trajectory of a particle. There is one notable exception to this, and it is important enough to look at in detail.

The geodesic equation can be written for the 'lowered' components of p as follows

$$p^{\alpha} p_{\beta;\alpha} = 0, \tag{7.25}$$

or

$$p^{\alpha} p_{\beta,\alpha} - \Gamma^{\gamma}{}_{\beta\alpha} p^{\alpha} p_{\gamma} = 0,$$

or

$$m \frac{\mathrm{d}p_{\beta}}{\mathrm{d}\tau} = \Gamma^{\gamma}{}_{\beta\alpha} p^{\alpha} p_{\gamma}. \tag{7.26}$$

Now, the right-hand side turns out to be simple

$$\begin{aligned} \Gamma^{\gamma}{}_{\alpha\beta} p^{\alpha} p_{\gamma} &= \frac{1}{2} g^{\gamma\nu} (g_{\nu\beta,\alpha} + g_{\nu\alpha,\beta} - g_{\alpha\beta,\nu}) p^{\alpha} p_{\gamma} \\ &= \frac{1}{2} (g_{\nu\beta,\alpha} + g_{\nu\alpha,\beta} - g_{\alpha\beta,\nu}) g^{\gamma\nu} p_{\gamma} p^{\alpha} \\ &= \frac{1}{2} (g_{\nu\beta,\alpha} + g_{\nu\alpha,\beta} - g_{\alpha\beta,\nu}) p^{\nu} p^{\alpha}. \end{aligned} \tag{7.27}$$

The product $p^\nu p^\alpha$ is symmetric on ν and α, while the first and third terms inside parentheses are, together, antisymmetric on ν and α. Therefore they cancel, leaving only the middle term

$$\Gamma^\gamma{}_{\beta\alpha} p^\alpha p_\gamma = \frac{1}{2} g_{\nu\alpha,\beta} p^\nu p^\alpha. \tag{7.28}$$

The geodesic equation can thus, in complete generality, be written

$$m \frac{\mathrm{d}p_\beta}{\mathrm{d}\tau} = \frac{1}{2} g_{\nu\alpha,\beta} p^\nu p^\alpha. \tag{7.29}$$

We therefore have the following important result: *if all the components $g_{\alpha\nu}$ are independent of x^β for some fixed index β, then p_β is a constant along any particle's trajectory*.

For instance, suppose we have a stationary (i.e. time-independent) gravitational field. Then a coordinate system can be found in which the metric components are time independent, and in that system p_0 is conserved. Therefore p_0 (or, really, $-p_0$) is usually called the 'energy' of the particle, *without* the qualification 'in this frame'. Notice that coordinates can also be found in which the same metric has time-dependent components: any time-dependent coordinate transformation from the 'nice' system will do this. In fact, most freely falling locally inertial systems are like this, since a freely falling particle sees a gravitational field that varies with its position, and therefore with time in its coordinate system. The frame in which the metric components are stationary is special, and is the usual 'laboratory frame' on Earth. Therefore p_0 in this frame is related to the usual energy defined in the lab, and includes the particle's gravitational potential energy, as we shall now show. Consider the equation

$$\begin{aligned}\vec{p} \cdot \vec{p} &= -m^2 = g_{\alpha\beta} p^\alpha p^\beta \\ &= -(1+2\phi)(p^0)^2 + (1-2\phi)[(p^x)^2 + (p^y)^2 + (p^z)^2],\end{aligned} \tag{7.30}$$

where we have used the metric, Eq. (7.8). This can be solved to give

$$(p^0)^2 = [m^2 + (1-2\phi)(p^2)](1+2\phi)^{-1}, \tag{7.31}$$

where, for shorthand, we denote by $\boldsymbol{p}^2$ the sum $(p^x)^2 + (p^y)^2 + (p^z)^2$. Keeping within the approximation $|\phi| \ll 1$, $|\boldsymbol{p}| \ll m$, we can simplify this to

$$(p^0)^2 \approx m^2(1 - 2\phi + \boldsymbol{p}^2/m^2)$$

or

$$p^0 \approx m(1 - \phi + \boldsymbol{p}^2/2m^2). \tag{7.32}$$

Now we lower the index and get

$$p^0 = g_{0\alpha}p^\alpha = g_{00}p^0 = -(1 + 2\phi)p^0, \tag{7.33}$$

$$-p_0 \approx m(1 + \phi + \boldsymbol{p}^2/2m^2) = m + m\phi + \boldsymbol{p}^2/2m. \tag{7.34}$$

The first term is the rest mass of the particle. The second and third are the Newtonian pieces of its energy: gravitational potential energy and kinetic energy. This means that the constancy of p_0 along a particle's trajectory generalizes the Newtonian concept of a conserved energy.

Notice that a *general* gravitational field will not be stationary in *any* frame,[2] so no conserved energy can be defined.

In a similar manner, if a metric is axially symmetric, then coordinates can be found in which $g_{\alpha\beta}$ is independent of the angle ψ around the axis. Then p_ψ will be conserved. This is the particle's angular momentum. In the nonrelativistic limit we have

$$p_\psi = g_{\psi\psi}p^\psi \approx g_{\psi\psi}m\,\mathrm{d}\psi/\mathrm{d}t \approx mg_{\psi\psi}\Omega, \tag{7.35}$$

where Ω is the angular velocity of the particle. Now, for a nearly flat metric we have

$$g_{\psi\psi} = \vec{e}_\psi \cdot \vec{e}_\psi \approx r^2 \tag{7.36}$$

in cylindrical coordinates (r, ψ, z) so that the conserved quantity is

$$p_\psi \approx mr^2\Omega. \tag{7.37}$$

This is the usual Newtonian definition of angular momentum.

So much for conservation laws of particle motion. Similar considerations apply to fluids, since they are just large collections of particles. But the situation with regard to the total mass and momentum of a self-gravitating system is more complicated. It turns out that an isolated system's mass and momentum *are* conserved, but we must postpone any discussion of this until we see how they are defined.

7.5 Further reading

The question of how curvature and physics fit together is discussed in more detail by Geroch (1978). Conserved quantities are discussed in detail in any of the advanced texts. The material in this chapter is preparation for the theory of quantum fields in a fixed curved spacetime. See Birrell and Davies (1984) and Wald (1994). This in turn leads to one of the most active areas of gravitation research today, the quantization of general relativity. While we will not treat this area in this book, readers in work that approaches this subject from the starting point of classical general relativity (as contrasted with approaching it from the starting point of string theory) may wish to look at Rovelli (2004) Bojowald (2005), and Thiemann (2007).

7.6 Exercises

If Eq. (7.3) were the correct generalization of Eq. (7.1) to a curved spacetime, how would you interpret it? What would happen to the number of particles in a comoving volume of the fluid, as time evolves? In principle, can we distinguish experimentally between Eqs. (7.2) and (7.3)?

To first order in ϕ, compute $g^{\alpha\beta}$ for Eq. (7.8).

Calculate all the Christoffel symbols for the metric given by Eq. (7.8), to first order in ϕ. Assume ϕ is a general function of t, x, y, and z.

Verify that the results, Eqs. (7.15) and (7.24), depended only on g_{00}: the form of g_{xx} doesn't affect them, as long as it is $1 + 0(\phi)$.

a) For a perfect fluid, verify that the spatial components of Eq. (7.6) in the Newtonian limit reduce to

$$v_{,t} + (v \cdot \nabla)v + \nabla p/\rho + \nabla\phi = 0 \tag{7.38}$$

for the metric, Eq. (7.8). This is known as Euler's equation for nonrelativistic fluid flow in a gravitational field. You will need to use Eq. (7.2) to get this

result.
(b) Examine the time-component of Eq. (7.6) under the same assumptions, and interpret each term.
(c) Eq. (7.38) implies that a static fluid (v = 0) in a static Newtonian gravitational field obeys the equation of hydrostatic equilibrium

$$\nabla p + \rho \nabla \phi = 0. \tag{7.39}$$

A metric tensor is said to be static if there exist coordinates in which $\vec{e}_0$ is timelike, $g_{i0} = 0$, and $g_{\alpha\beta,0} = 0$. Deduce from Eq. (7.6) that a static fluid ($U^i = 0$, $p_{,0} = 0$, etc.) obeys the relativistic equation of hydrostatic equilibrium

$$p_{,i} + (\rho + p)\left[\tfrac{1}{2}\ln(-g_{00})\right]_{,i} = 0. \tag{7.40}$$

(d) This suggests that, at least for static situations, there is a close relation between g_{00} and $-\exp(2\phi)$, where ϕ is the Newtonian potential for a similar physical situation. Show that Eq. (7.8) and Exer. 4 are consistent with this.
Deduce Eq. (7.25) from Eq. (7.10).
Consider the following four different metrics, as given by their line elements:
(i) $ds^2 = -dt^2 + dx^2 + dy^2 + dz^2$;
(ii) $ds^2 = -(1 - 2M/r)\, dt^2 + (1 - 2M/r)^{-1}\, dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2)$, where M is a constant;
(iii)

$$ds^2 = -\frac{\Delta - a^2\sin^2\theta}{\rho^2}dt^2 - 2a\frac{2Mr\sin^2\theta}{\rho^2}dt\, d\phi + \frac{(r^2+a^2)^2 - a^2\Delta\sin^2\theta}{\rho^2}\sin^2\theta\, d\phi^2 + \frac{\rho^2}{\Delta}dr^2 + \rho^2\, d\theta^2,$$

where M and a are constants and we have introduced the shorthand notation $\Delta = r^2 - 2Mr + a^2$, $\rho^2 = r^2 + a^2\cos^2\theta$;
(iv) $ds^2 = -dt^2 + R^2(t)\,[(1 - kr^2)^{-1}dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2)]$, where k is a constant and $R(t)$ is an arbitrary function of t alone.

The first one should be familiar by now. We shall encounter the other three in later chapters. Their names are, respectively, the Schwarzschild, Kerr, and Robertson–Walker metrics.

(a) For each metric find as many conserved components ρ_α of a freely falling particle's four momentum as possible.

(b) Use the result of Exer. 28, § 6.9 to put (i) in the form

$$(\mathrm{i}')\ \mathrm{d}s^2 = -\mathrm{d}t^2 + \mathrm{d}r^2 + r^2(\mathrm{d}\theta^2 + \sin^2\theta\ \mathrm{d}\phi^2).$$

From this, argue that (ii) and (iv) are spherically symmetric. Does this increase the number of conserved components p_α?

(c) It can be shown that for (i′) and (ii)–(iv), a geodesic that begins with $\theta = \pi/2$ and $p^\theta = 0$ – *i.e.* one which begins tangent to the equatorial plane – always has $\theta = \pi/2$ and $p^\theta = 0$. For cases (i′), (ii), and (iii), use the equation $\vec{p} \cdot \vec{p} = -m^2$ to solve for p^r in terms of m, other conserved quantities, and known functions of position.

(d) For (iv), spherical symmetry implies that if a geodesic begins with $p^\theta = p^\phi = 0$, these remain zero. Use this to show from Eq. (7.29) that when $k = 0$, p_r is a conserved quantity.

Suppose that in some coordinate system the components of the metric $g_{\alpha\beta}$ are independent of some coordinate x^μ.

(a) Show that the conservation law $T^\nu{}_{\mu;\nu} = 0$ for *any* stress–energy tensor becomes

$$\frac{1}{\sqrt{-g}}(\sqrt{-g}T^\nu{}_\mu)_{,\nu} = 0. \qquad (7.41)$$

(b) Suppose that in these coordinates $T^{\alpha\beta} \neq 0$ only in some bounded region of each spacelike hypersurface $x^0 = \text{const}$. Show that Eq. (7.41) implies

$$\int_{x^0=\text{const.}} T^\nu{}_\mu \sqrt{-g}\, n_\nu\, \mathrm{d}^3x$$

is independent of x^0, if n_ν is the unit normal to the hypersurface. This is the generalization to continua of the conservation law stated after Eq. (7.29).

(c) Consider flat Minkowski space in a global inertial frame with spherical polar coordinates (t, r, θ, ϕ). Show from (b) that

$$J = \int_{t=\text{const.}} T^0{}_\phi r^2 \sin\theta\, \mathrm{d}r\, \mathrm{d}\theta\, \mathrm{d}\phi \qquad (7.42)$$

is independent of t. This is the total angular momentum of the system.

(d) Express the integral in (c) in terms of the components of $T^{\alpha\beta}$ on the Cartesian basis (t, x, y, z), showing that

$$J = \int (xT^{y0} - yT^{x0})\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z. \tag{7.43}$$

This is the continuum version of the nonrelativistic expression $(\boldsymbol{r} \times \boldsymbol{p})_z$ for a particle's angular momentum about the z axis.

a) Find the components of the Riemann tensor $R_{\alpha\beta\mu\nu}$ for the metric, Eq. (7.8), to first order in ϕ.

(b) Show that the equation of geodesic deviation, Eq. (6.87), implies (to lowest order in ϕ and velocities)

$$\frac{\mathrm{d}^2\xi^i}{\mathrm{d}t^2} = -\phi_{,ij}\xi^j. \tag{7.44}$$

(c) Interpret this equation when the geodesics are world lines of freely falling particles which begin from rest at nearby points in a Newtonian gravitational field.

) Show that if a vector field ξ^α satisfies *Killing's equation*

$$\nabla_\alpha \xi_\beta + \nabla_\beta \xi_\alpha = 0, \tag{7.45}$$

then along a geodesic, $p^\alpha\xi_\alpha$ = const. This is a coordinate-invariant way of characterizing the conservation law we deduced from Eq. (7.29). We only have to know whether a metric admits Killing fields.

(b) Find ten Killing fields of Minkowski spacetime.

(c) Show that if $\vec{\xi}$ and $\vec{\eta}$ are Killing fields, then so is $\alpha\vec{\xi} + \beta\,\vec{\eta}$ for *constant* α and β.

(d) Show that Lorentz transformations of the fields in (b) simply produce linear combinations as in (c).

(e) If you did Exer. 7, use the results of Exer. 7(a) to find Killing vectors of metrics (ii)–(iv).

It is more common to define the WEP without reference to a curved spacetime, but just to say that all particles fall at the same rate in a gravitational field, independent of their mass and composition. But the Einstein Equivalence Principle (Postulate IV′) is normally taken to imply that gravity can be represented by spacetime curvature, so we shall simply start with the assumption that we have a curved spacetime.

It is easy to see that there is generally no coordinate system which makes a given metric time independent. The metric has ten independent components (same as a 4 × 4 symmetric matrix), while a change of coordinates enables us to introduce only four degrees of freedom to change the components (these are the four functions $x^{\bar{\alpha}}(x^{\mu})$). It is a special metric indeed if all ten components can be made time independent this way.

9

Gravitational radiation

9.1 The propagation of gravitational waves

It may happen that in a region of spacetime the gravitational field is weak but not stationary. This can happen far from a fully relativistic source undergoing rapid changes that took place long enough ago for the disturbances produced by the changes to reach the distant region under consideration. We shall study this problem by using the weak-field equations developed in the last chapter, but first we study the solutions of the homogeneous system of equations that we excluded from the Newtonian treatment in § 8.4. The Einstein equations Eq. (8.42), in vacuum ($T^{\mu\nu} = 0$) far outside the source of the field, are

$$\left(-\frac{\partial^2}{\partial t^2} + \nabla^2\right)\bar{h}^{\alpha\beta} = 0. \tag{9.1}$$

In this chapter we do not neglect $\partial^2/\partial t^2$. Eq. (9.1) is called the three-dimensional wave equation. We shall show that it has a (complex) solution of the form

$$\bar{h}^{\alpha\beta} = A^{\alpha\beta}\ \exp\left(\mathrm{i}k_\alpha x^\alpha\right), \tag{9.2}$$

where $\{k_\alpha\}$ are the (real) constant components of some one-form and $\{A^{\alpha\beta}\}$ the (complex) constant components of some tensor. (In the end we shall take the real part of any complex solutions.) Eq. (9.1) can be written as

$$\eta^{\mu\nu}\bar{h}^{\alpha\beta}{}_{,\mu\nu} = 0, \tag{9.3}$$

and, from Eq. (9.2), we have

$$\bar{h}^{\alpha\beta}{}_{,\mu} = \mathrm{i}k_\mu\bar{h}^{\alpha\beta}. \tag{9.4}$$

Therefore, Eq. (9.3) becomes

$$\eta^{\mu\nu}\bar{h}^{\alpha\beta}{}_{,\mu\nu} = -\eta^{\mu\nu}k_\mu k_\nu\bar{h}^{\alpha\beta} = 0.$$

This can vanish only if

$$\eta^{\mu\nu}k_\mu k_\nu = k^\nu k_\nu = 0. \tag{9.5}$$

So Eq. (9.2) gives a solution to Eq. (9.1) if k_α is a *null* one-form or, equivalently, if the associated four-vector k^α is *null*, *i.e.* tangent to the world line of a photon. (Recall that we raise and lower indices with the flat-space metric tensor $\eta^{\mu\nu}$, so k^α is a Minkowski null vector.) Eq. (9.2) describes a wavelike solution. The value of $\bar{h}^{\alpha\beta}$ is constant on a hypersurface on which $k_\alpha x^\alpha$ is constant:

$$k_\alpha x^\alpha = k_0 t + \boldsymbol{k}\cdot\boldsymbol{x} = \text{const.}, \tag{9.6}$$

where $\boldsymbol{k}$ refers to $\{k^i\}$. It is conventional to refer to k^0 as ω, which is called the frequency of the wave:

$$\vec{k} \to (\omega, \boldsymbol{k}) \tag{9.7}$$

is the time–space decomposition of $\vec{k}$. Imagine a photon moving in the direction of the null vector $\vec{k}$. It travels on a curve

$$x^\mu(\lambda) = k^\mu\lambda + l^\mu, \tag{9.8}$$

where λ is a parameter and l^μ is a constant vector (the photon's position at λ = 0). From Eqs. (9.8) and (9.5), we find

$$k_\mu x^\mu(\lambda) = k_\mu l^\mu = \text{const.} \tag{9.9}$$

Comparing this with Eq. (9.6), we see that the photon travels with the gravitational wave, staying forever at the same phase. We express this by saying that the wave itself travels at the speed of light, and $\vec{k}$ is its direction of travel. The nullity of $\vec{k}$ implies

$$\omega^2 = |\boldsymbol{k}|^2, \tag{9.10}$$

which is referred to as the dispersion relation for the wave. Readers familiar with wave theory will immediately see from Eq. (9.10) that the wave's phase velocity is one, as is its group velocity, and that there is no dispersion.

The Einstein equations only assume the simple form, Eq. (9.1), if we impose the gauge condition

$$\bar{h}^{\alpha\beta}{}_{,\beta} = 0, \tag{9.11}$$

the consequences of which we must therefore consider. From Eq. (9.4), we find

(9.12)

$$A^{\alpha\beta}k_\beta = 0, \tag{9.12}$$

which is a restriction on $A^{\alpha\beta}$: it must be orthogonal to $\vec{k}$

The solution $A^{\alpha\beta}$ exp$(ik_\mu x^\mu)$ is called a *plane wave*. (Of course, in physical applications, we use only the real part of this expression, allowing $A^{\alpha\beta}$ to be complex.) By the theorems of Fourier analysis, *any* solution of Eqs. (9.1) and (9.11) is a superposition of plane wave solutions (see Exer. 3, § 9.7).

The transverse-traceless gauge

We so far have only one constraint, Eq. (9.12), on the amplitude $A^{\alpha\beta}$, but we can use our gauge freedom to restrict it further. Recall from Eq. (8.38) that we can change the gauge while remaining within the Lorentz class of gauges using any vector solving

$$\left(-\frac{\partial^2}{\partial t^2} + \nabla^2\right)\xi_\alpha = 0. \tag{9.13}$$

Let us choose a solution

$$\xi_\alpha = B_\alpha \exp(\mathrm{i}k_\mu x^\mu), \tag{9.14}$$

where B_α is a constant and k^μ is the same null vector as for our wave solution. This produces a change in $h^{\alpha\beta}$, given by Eq. (8.24),

$$h_{\alpha\beta}^{(\mathrm{NEW})} = h_{\alpha\beta}^{(\mathrm{OLD})} - \xi_{\alpha,\beta} - \xi_{\beta,\alpha} \tag{9.15}$$

and a consequent change in $\bar{h}_{\alpha\beta}$, given by Eq. (8.34),

$$\bar{h}_{\alpha\beta}^{(\mathrm{NEW})} = \bar{h}_{\alpha\beta}^{(\mathrm{OLD})} - \xi_{\alpha,\beta} - \xi_{\beta,\alpha} + \eta_{\alpha\beta}\xi^\mu{}_{,\mu}. \tag{9.16}$$

Using Eq. (9.14) and dividing out the exponential factor common to all terms gives

$$A_{\alpha\beta}^{(\mathrm{NEW})} = A_{\alpha\beta}^{(\mathrm{OLD})} - \mathrm{i}\,B_\alpha k_\beta - \mathrm{i}\,B_\beta k_\alpha + \mathrm{i}\,\eta_{\alpha\beta}B^\mu k_\mu. \tag{9.17}$$

In Exer. 5, § 9.7, it is shown that B_α can be chosen to impose two further restrictions on $A_{\alpha\beta}^{(\mathrm{NEW})}$:

$$A^\alpha{}_\alpha = 0 \tag{9.18}$$

and

$$A_{\alpha\beta} U^{\beta} = 0, \tag{9.19}$$

where $\vec{U}$ is some fixed four-velocity, *i.e.* any constant timelike unit vector we wish to choose. Eqs. (9.12), (9.18), and (9.19) together are called the *transverse–traceless* (TT) gauge conditions. (The word 'traceless' refers to Eq. (9.18); 'transverse' will be explained below.) We have now used up all our gauge freedom, so any remaining independent components of $A_{\alpha\beta}$ must be physically important. Notice, by the way, that the trace condition, Eq. (9.18), implies (see Eq. (8.29))

$$\bar{h}^{\mathrm{TT}}_{\alpha\beta} = h^{\mathrm{TT}}_{\alpha\beta}. \tag{9.20}$$

Let us go to a Lorentz frame for the background Minkowski spacetime (i.e. make a background Lorentz transformation), in which the vector $\vec{U}$ upon which we have based the TT gauge is the time basis vector $U^{\beta} = \delta^{\beta}0$. Then Eq. (9.19) implies $A_{\alpha 0} = 0$ for all α. In this frame, let us orient our spatial coordinate axes so that the wave is traveling in the z direction, $\vec{k} \to (\omega, 0, 0, \omega)$. Then, with Eq. (9.19), Eq. (9.12) implies $A_{\alpha z} = 0$ for all α. (This is the origin of the adjective 'transverse' for the gauge: $A_{\mu\nu}$ is 'across' the direction of propagation $\vec{e}_z$.) These two restrictions mean that only A_{xx}, A_{yy}, and $A_{xy} = A_{yx}$ are nonzero. Moreover, the trace condition, Eq. (9.18), implies $A_{xx} = -A_{yy}$. In matrix form, we therefore have in this specially chosen frame

$$(A^{\mathrm{TT}}_{\alpha\beta}) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & A_{xx} & A_{xy} & 0 \\ 0 & A_{xy} & -Axx & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{9.21}$$

There are only *two* independent constants, A^{TT}_{xx} and A^{TT}_{xy}. What is their physical significance?

The effect of waves on free particles

As we remarked earlier, any wave is a superposition of plane waves; if the wave travels in the z direction, we can put all the plane waves in the form of Eq. (9.21), so that any wave has only the two independent components h^{TT}_{xx} and h^{TT}_{xy}.

Consider a situation in which a particle initially in a wave-free region of spacetime encounters a gravitational wave. Choose a background Lorentz frame in which the particle is initially at rest, and choose the TT gauge referred to in this frame (i.e. the four-velocity U^α in Eq. (9.19) is the initial four-velocity of the particle). A free particle obeys the geodesic equation, Eq. (7.9),

$$\frac{\mathrm{d}}{\mathrm{d}\tau}U^\alpha + \Gamma^\alpha{}_{\mu\nu}U^\mu U^\nu = 0. \tag{9.22}$$

Since the particle is initially at rest, the initial value of its acceleration is

$$(\mathrm{d}U^\alpha/\mathrm{d}\tau)_0 = -\Gamma^\alpha{}_{00} = -\tfrac{1}{2}\eta^{\alpha\beta}(h_{\beta 0,0} + h_{0\beta,0} - h_{00,\beta}). \tag{9.23}$$

But by Eq. (9.21), $h^{\mathrm{TT}}_{\beta 0}$ vanishes, so initially the acceleration vanishes. This means the particle will still be at rest a moment later, and then, by the same argument, the acceleration will still be zero a moment later. The result is that the particle remains at rest forever, regardless of the wave! However, being ‘at rest’ simply means remaining at a constant coordinate position, so we should not be too hasty in its interpretation. All we have discovered is that by choosing the TT gauge – which means making a particular adjustment in the ‘wiggles’ of our coordinates – we have found a coordinate system that stays attached to individual particles. This in itself has no invariant geometrical meaning.

To get a better measure of the effect of the wave, let us consider two nearby particles, one at the origin and another at $x = \varepsilon$, $y = z = 0$, both beginning at rest. Both then remain at these coordinate positions, and the proper distance between them is

$$\begin{aligned}\Delta l \equiv \int |\mathrm{d}s^2|^{1/2} &= \int |g_{\alpha\beta}\ \mathrm{d}x^\alpha\ \mathrm{d}x^\beta|^{1/2} \\ &= \int_0^\varepsilon |g_{xx}|^{1/2}\ \mathrm{d}x \approx |g_{xx}(x=0)|^{1/2}\varepsilon \\ &\approx [1 + \tfrac{1}{2}h^{\mathrm{TT}}_{xx}(x=0)]\varepsilon.\end{aligned} \tag{9.24}$$

Now, since h^{TT}_{xx} is not generally zero, the *proper* distance (as opposed to the coordinate distance) does change with time. This is an illustration of the difference between computing a coordinate-dependent number (the position of a particle) and a coordinate-independent number (the proper distance between two particles). The effect of the wave is unambiguously seen in the coordinate-independent number. The proper distance between two particles can be measured: we will discuss two ways of measuring it in the paragraph on 'Measuring the stretching of space' below. The physical effects of gravitational fields always show up in measurables.

Equation (9.24) tells us a lot. First, the change in the distance between two particles is proportional to their initial separation ε. Gravitational waves create a bigger distance change if the original distance is bigger. This is the reason that modern gravitational wave detectors, which we discuss below, are designed and built on huge scales, measuring changes in separations over many kilometers (for ground-based detectors) or millions of kilometers (in space). The second thing we learn from Eq. (9.24) is that the effect is small, proportional to h_{ij}^{TT}. We will see when we study the generation of waves below that these dimensionless components are typically 10^{-21} or smaller. So gravitational wave detectors have to sense relative distance changes of order one part in 10^{21}. This is the experimental challenge that was achieved for the first time in 2005, and improvements in sensitivity are continually being made.

Tidal accelerations: gravitational wave forces

Another approach to the same question of how gravitational waves affect free particles involves the equation of geodesic deviation, Eq. (6.87). This will lead us, in the following paragraph, to a way of understanding the action of gravitational waves as a tidal force on particles, whether they are free or not.

Consider again two freely falling particles, and set up the connecting vector ξ^α between them. If we were to work in a TT-coordinate system, as in the previous paragraph, then the fact that the particles remain at rest in the coordinates means that the components of $\vec{\xi}$ would remain constant; although correct, this would not be a helpful result since we have not associated the components of $\vec{\xi}$ in TT-coordinates with the result of any measurement. Instead we shall work in a coordinate system closely associated with measurements, the local inertial frame at the point of the first geodesic where $\vec{\xi}$ originates. In this frame, coordinate distances are proper distances, as long as we can neglect quadratic terms in the coordinates. That means that in these coordinates the components of $\vec{\xi}$ do indeed correspond to measurable proper distances if the geodesics are near enough to one another.

What is more, in this frame the second covariant derivative in Eq. (6.87) simplifies. It starts out as $\nabla_U \nabla_U \xi^\alpha$, where we are calling the tangent to the geodesic $\vec{U}$ here instead of $\vec{V}$. Now, the first derivative acting on $\vec{\xi}$ just gives $d\xi^\alpha/d\tau$. But the second derivative is a covariant one, and should contain not just $d/d\tau$ but also a term with a Christoffel symbol. But in this local inertial frame the Christoffel symbols all vanish at this point, so the second derivative is just an ordinary second derivative with respect to *tau*. The result is, again in the locally inertial frame,

$$\frac{\mathrm{d}^2}{\mathrm{d}\tau^2}\xi^\alpha = R^\alpha{}_{\mu\nu\beta}\, U^\mu U^\nu \xi^\beta, \tag{9.25}$$

where $\vec{U} = \mathrm{d}\bar{x}/\mathrm{d}\tau$ is the four-velocity of the two particles. In these coordinates the components of $\vec{U}$ are needed only to lowest (i.e. flat-space) order, since any corrections to U^α that depend on $h_{\mu\nu}$ will give terms second order in $h_{\mu\nu}$ in the above equation (because $R^\alpha{}_{\mu\nu\beta}$ is already first order in $h_{\mu\nu}$). Therefore, $\vec{U} \to (1, 0, 0, 0)$ and, initially, $\vec{\xi} \to (0, \varepsilon, 0, 0)$. Then, to first order in $h_{\mu\nu}$, Eq. (9.25) reduces to

$$\frac{\mathrm{d}^2}{\mathrm{d}\tau^2}\xi^\alpha = \frac{\partial^2}{\partial t^2}\xi^\alpha = \varepsilon R^\alpha{}_{00x} = -\varepsilon R^\alpha{}_{0x0}. \tag{9.26}$$

This is the fundamental result, which shows that the Riemann tensor is locally measurable by simply watching the proper distance changes between nearby geodesics.

Now, the Riemann tensor is itself gauge invariant, so its components do not depend on the choice we made between a local inertial frame and the TT coordinates. It follows also that the left-hand side of Eq. (9.26) must have an interpretation independent of the coordinate gauge. We identify ξ^α as the *proper lengths* of the components of the connecting vector $\vec{\xi}$, in other words the proper distances along the four coordinate directions over the coordinate intervals spanned by the vector. With this interpretation, we free ourselves from the choice of gauge and arrive at a gauge-invariant interpretation of the whole of Eq. (9.26).

Just to emphasize that we have restored gauge freedom to this equation, let us write the Riemann tensor components in terms of the components of the metric in TT gauge. This is possible, since the Riemann components are gauge-invariant. And it is desirable, since these components are particularly simple in the TT gauge. It is not hard to use Eq. (8.25) to show that, for a wave traveling in the z-direction, the components are

$$\left.\begin{aligned} R^{x}{}_{0x0} &= R_{x0x0} = -\tfrac{1}{2} h^{\mathrm{TT}}_{xx,00}, \\ R^{y}{}_{0x0} &= R_{y0x0} = -\tfrac{1}{2} h^{\mathrm{TT}}_{xy,00}, \\ R^{y}{}_{0y0} &= R_{y0y0} = -\tfrac{1}{2} h^{\mathrm{TT}}_{yy,00} = -R^{x}{}_{0x0}, \end{aligned}\right\} \tag{9.27}$$

with all other independent components vanishing. This means, for example, that two particles initially separated in the x direction have a separation vector $\vec{\xi}$ whose components' proper lengths obey

$$\frac{\partial^2}{\partial t^2}\xi^x = \tfrac{1}{2}\varepsilon \frac{\partial^2}{\partial t^2} h^{\mathrm{TT}}_{xx}, \qquad \frac{\partial^2}{\partial t^2}\xi^y = \tfrac{1}{2}\varepsilon \frac{\partial^2}{\partial t^2} h^{\mathrm{TT}}_{xy}. \tag{9.28a}$$

This is clearly consistent with Eq. (9.24). Similarly, two particles initially separated by ε in the y direction obey

$$\begin{aligned} \frac{\partial^2}{\partial t^2}\xi^y &= \tfrac{1}{2}\varepsilon \frac{\partial^2}{\partial t^2} h^{\mathrm{TT}}_{yy} = -\tfrac{1}{2}\varepsilon \frac{\partial^2}{\partial t^2} h^{\mathrm{TT}}_{xx}, \\ \frac{\partial^2}{\partial t^2}\xi^x &= \tfrac{1}{2}\varepsilon \frac{\partial^2}{\partial t^2} h^{\mathrm{TT}}_{xy}. \end{aligned} \tag{9.28b}$$

Remember, from Eq. (9.21), that $h^{\mathrm{TT}}_{yy} = -h^{\mathrm{TT}}_{xx}$.

Measuring the stretching of space

The action of gravitational waves is sometimes characterized as a stretching of space. Eq. (9.24) makes it clear what this means: as the wave passes through, the proper separations of free objects that are simply sitting at rest change with time. However, students of general relativity sometimes find this concept confusing. A frequent question is, if space is stretched, why is a ruler (which consists, after all, mostly of empty space with a few electrons and nuclei scattered through it) not also stretched, so that the stretching is not measurable by the ruler? The answer to this question lies not in Eq. (9.24) but in the geodesic deviation equation, Eq. (9.26).

Although the two formulations of the action of a gravitational wave, Eqs. (9.24) and (9.26), are essentially equivalent for free particles, the second one is far more useful and instructive when we consider the behavior of particles that have other forces acting on them as well. The first formulation is the complete solution for the relative motion of particles that are freely falling, but it gives no way of including other forces. The second formulation is not a *solution* but a *differential equation*. It shows the acceleration of one particle (let's call it B), induced by the wave, as measured in a freely falling local inertial frame that initially coincides with the motion of the other particle (let's call it A). It says that, in this local frame, the wave acts just like a force pushing on B. This is called the *tidal force* associated with the wave. This force depends on the separation $\vec{\xi}$. From Eq. (9.25), it is clear that the acceleration resulting from this effective force is

$$\frac{\partial^2}{\partial t^2}\xi^i = -R^i{}_{0j0}\xi^j. \tag{9.29}$$

Now, if particle B has another force on it as well, say $\vec{F}_\mathrm{B}$, then to get the complete motion of B we must simply solve Newton's second law with all forces included. This means solving the differential equation

$$\frac{\partial^2}{\partial t^2}\xi^i = -R^i{}_{0j0}\xi^j + \frac{1}{m_B}F^i_B, \tag{9.30}$$

where m_B is the mass of particle B. Indeed, if particle A also has a force $\vec{F}_A$ on it, then it will not remain at rest in the local inertial frame. But we can still make measurements in that frame, and in this case the separation of the two particles will obey

$$\frac{\partial^2}{\partial t^2}\xi^i = -R^i{}_{0j0}\xi^j + \frac{1}{m_B}F^i_B - \frac{1}{m_A}F^i_A. \tag{9.31}$$

This equation allows us to treat material systems acted on by gravitational waves. The continuum version of it can be used to understand how a 'bar' detector of gravitational waves, which we will encounter below, reacts to an incident wave. And it allows us to answer the question of what happens to a ruler when the wave hits. Since the atoms in the ruler are not free, but instead are acted upon by electric forces from nearby atoms, the ruler will stretch by an amount that depends on how strong the tidal gravitational forces are compared to the internal binding forces. Now, gravitational forces are very weak compared to electric forces, so in practice the ruler does not stretch at all. In this way the ruler can be used to measure the tidal displacement of nearby free particles, in other words to measure the 'stretching of space'.

There are other ways of measuring the stretching. One of the most important in practice is to send photons back and forth between the free particles and measure the changes in the light–travel time between them. This is the principle of the laser interferometer gravitational wave detector, and we will discuss it in some detail below.

Polarization of gravitational waves

Eqs. (9.28a) and (9.28b) form the foundation of the definition of the *polarization* of a gravitational wave. Consider a ring of particles initially at rest in the $x - y$ plane, as in Fig. 9.1(*a*). Suppose a wave has $h^{TT}_{xx} \neq 0$, $h^{TT}_{xy} = 0$. Then the particles will be moved (in terms of proper distance relative to the one in the center) in the way shown in Fig. 9.1(*b*), as the wave oscillates and $h^{TT}_{xx} = -h^{TT}_{yy}$ changes sign. If, instead, the wave had $h^{TT}_{xy} \neq 0$ but $h^{TT}_{xx} = h^{TT}_{yy} = 0$, then the picture would distort as in Fig. 9.1(*c*). Since h^{TT}_{xy} and h^{TT}_{xx} are independent, (*b*) and (*c*) provide a pictorial representation for two different linear polarizations. Notice that the two states are simply rotated 45° relative to one another. This contrasts with the two polarization states of an electromagnetic wave, which are 90° to each other. As Exer. 16, § 9.7, shows, this pattern of polarization is due to the fact that gravity is represented by the second-rank symmetric tensor $h_{\mu\nu}$. (By contrast, electromagnetism is represented by the vector potential A^{μ} of Exer. 11, § 8.6.)

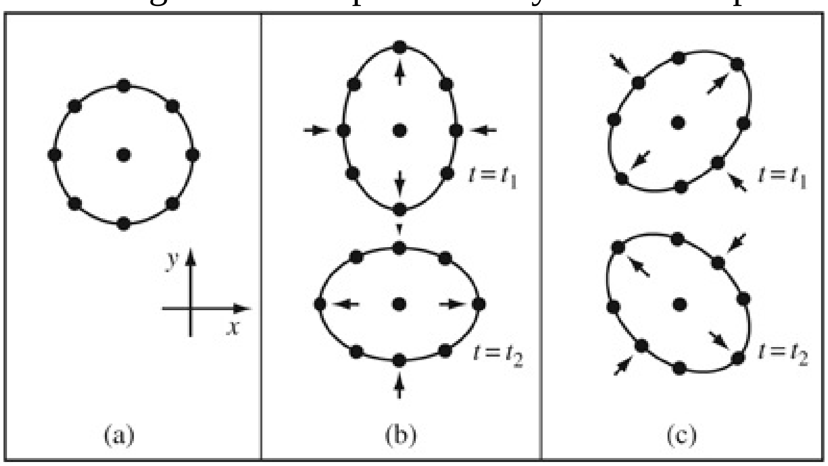


Figure 9.1 (a) A circle of free particles before a wave traveling in the *z* direction reaches them. (b) Distortions of the circle produced by a wave with the '+' polarization. The two pictures represent the same wave at phases separated by 180°. Particles are positioned according to their proper distances from one another. (c) As (b) for the '×' polarization.

An exact plane wave

Although all waves that we can expect to detect on Earth are so weak that linearized theory ought to describe them very accurately, it is interesting to see if the linear plane wave corresponds to some exact solution of the nonlinear equations that has similar properties. We shall briefly derive such a solution.

Suppose the wave is to travel in the z direction. By analogy with Eq. (9.2), we might hope to find a solution that depends only on $u := t - z.$

This suggests using u as a coordinate, as we did in Exer. 34, § 3.10 (with x replaced by z). In flat space it is natural, then, to define a complementary null coordinate $v = t + z,$

so that the line element of flat spacetime becomes

$$ds^2 = -du\, dv + dx^2 + dy^2.$$

Now, we have seen that the linear wave affects only proper distances perpendicular to its motion, *i.e.* in the $x - y$ coordinate plane. So let us look for a nonlinear generalization of this, *i.e.* for a solution with the metric

$$ds^2 = -du\, dv + f^2(u)\, dx^2 + g^2(u)\, dy^2,$$

where f and g are functions to be determined by Einstein's equations. It is a straightforward calculation to discover that the only nonvanishing Christoffel symbols and Riemann tensor components are

$$\begin{aligned}
\Gamma^x{}_{xu} &= \dot{f}/f, \quad \Gamma^y{}_{yu} = \dot{g}/g, \\
\Gamma^v{}_{xx} &= 2\dot{f}/f, \quad \Gamma^v{}_{yy} = 2\dot{g}/g, \\
R^x{}_{uxu} &= -\ddot{f}/f, \quad R^y{}_{uyu} = -\ddot{g}/g,
\end{aligned}$$

and others obtainable by symmetries. Here, dots denote derivatives with respect to u. The only vacuum field equation then becomes

$$\ddot{f}/f + \ddot{g}/g = 0. \tag{9.32}$$

We can therefore prescribe an arbitrary function $g(u)$ and solve this equation for $f(u)$. This is the same freedom as we had in the linear case, where Eq. (9.2) can be multiplied by an arbitrary $f(k_z)$ and integrated over k_z to give the Fourier representation of an arbitrary function of $(z - t)$. In fact, if g is nearly 1, $g \approx 1 + \varepsilon(u),$

so we are near the linear case, then Eq. (9.32) has a solution

$$f \approx 1 - \varepsilon(a).$$

This is just the linear wave in Eq. (9.21), with plane polarization with $h_{xy} = 0$, *i.e.* the polarization shown in Fig. 9.1(b). Moreover, it is easy to see that the geodesic equation implies, in the nonlinear case, that a particle initially at rest on our coordinates remains at rest. We have, therefore, a simple nonlinear solution corresponding to our approximate linear one.

This solution is one of a class called plane-fronted waves with parallel rays. See Ehlers and Kundt (1962), § 2.5, and Stephani, *et al.* (2003), § 24.5.

Geometrical optics: waves in a curved spacetime

In this chapter we have made the simplifying assumption that our gravitational waves are the only gravitational field, that they are perturbations of flat spacetime, starting from Eq. (8.12). We found that they behave like a wave field moving at the speed of light in special relativity. But the real universe contains other gravitational fields, and gravitational waves have to make their way to our detectors through the fields of stars, galaxies, and the universe as a whole. How do they move?

The answer comes from studying waves as perturbations of a curved metric, of the form $g_{\alpha\beta} + h_{\alpha\beta}$, where **g** could be the metric created by any combination of sources of gravity. The computation of the dynamical equation governing **h** is very similar to the one we went through at the beginning of this chapter, but the mathematics of curved spacetime must be used. We won't go into the details here, but it is important to understand qualitatively that the results are very similar to the results we would get for electromagnetic waves traveling through complicated media.

If the waves have short wavelength, then they basically follow a null geodesic, and they parallel-transport their polarization tensor. This is exactly the same as for electromagnetic waves, so that photons and gravitational waves leaving the same source at the same time will continue to travel together through the universe, provided they move through vacuum. For this *geometrical optics* approximation to hold, the wavelength has to be short in two ways. It must be short compared to the typical curvature scale, so that the wave is merely a ripple on a smoothly curved background spacetime; and its period must be short compared to the timescale on which the background gravitational fields might change. If nearby null geodesics converge, then gravitational and electromagnetic waves traveling on them will be focussed and become stronger. This is called *gravitational lensing*, and we will see an example of it in Ch. 11.

Gravitational waves do not always keep step with their electromagnetic counterparts. Electromagnetic waves are strongly affected by ordinary matter, so that if their null geodesic passes through matter, they can suffer additional lensing, scattering, or absorption. Gravitational waves are hardly disturbed by matter at all. They follow the null geodesics even through matter. The reason is the weakness of their interaction with matter, as we saw in Eq. (9.24). If the wave amplitudes h_{ij}^{TT} are small, then their effect on any matter they pass through is also small, and the back-effect of the matter on them will be of the same order of smallness. Gravitational waves are therefore highly prized carriers of information from distant regions of the universe: we can in principle use them to 'see' into the centers of supernova explosions, through obscuring dust clouds, or right back to the first fractions of a second after the Big Bang.

9.2 The detection of gravitational waves

General considerations

The great progress that astronomy has made since about 1960 is due largely to the fact that technology has permitted astronomers to begin to observe in many different parts of the electromagnetic spectrum. Because they were restricted to observing visible light, the astronomers of the 1940s could have had no inkling of such diverse and exciting phenomena as quasars, pulsars, black holes in X-ray binaries, giant black holes in galactic centers, gamma-ray bursts, and the cosmic microwave background radiation. As technology has progressed, each new wavelength region has revealed unexpected and important information. Most regions of the electromagnetic spectrum have now been explored at some level of sensitivity, but there is another spectrum which is as yet completely untouched: the gravitational wave spectrum.

As we shall see in § 9.5 below, nearly all astrophysical phenomena emit gravitational waves, and the most violent ones (which are of course among the most interesting ones!) give off radiation in copious amounts. In some situations, gravitational radiation carries information that no electromagnetic radiation can give us. For example, gravitational waves come to us direct from the heart of supernova explosions; the electromagnetic radiation from the same region is scattered countless times by the dense material surrounding the explosion, taking days to eventually make its way out, and in the process losing most of the detailed information it might carry about the explosion. As another example, gravitational waves from the Big Bang originated when the universe was perhaps only 10^{-25} s old; they are our earliest messengers from the beginning of our universe, and they should carry the imprint of unknown physics at energies far higher than anything we can hope to reach in accelerators on the Earth.

Beyond what we can predict, we can be virtually certain that the gravitational-wave spectrum has surprises for us; clues to phenomena we never suspected. Astronomers know that only 4% of the mass-energy of the universe is in charged particles that can emit or receive electromagnetic waves; the remaining 96% cannot radiate electromagnetically but it nevertheless couples to gravity, and some of it could turn out to radiate gravitational waves. It is not surprising, therefore, that considerable effort has been devoted to the development of sensitive gravitational-wave antennas.

The technical difficulties involved in the detection of gravitational radiation are enormous, because the amplitudes of the metric perturbations $h_{\mu\nu}$ that can be expected from distant sources are so small (see §§ 9.3 and 9.5 below). This is an area in which rapid advances are being made in a complex interplay between advancing technology, large investments by scientific funding agencies, and new astrophysical discoveries. This chapter reflects the situation in 2008, when sensitive detectors are in operation, even more sensitive ones are planned, but no direct detections have yet been made. The student who wants to get updated should consult the scientific literature referred to in the bibliography, § 9.6.

Purpose-built detectors are of two types: bars and interferometers.

- *Resonant mass detectors.* Also known as 'bar detectors', these are solid masses that respond to incident gravitational waves by going into vibration. The first purpose-built detectors were of this kind (Weber 1961). We shall study them below, because they can teach us a lot. But they are being phased out, because interferometers (below) have reached a better sensitivity.
- *Laser interferometers.* These detectors use highly stable laser light to monitor the proper distances between free masses; when a gravitational wave comes by, these distances will change – as we saw earlier. The principle is illustrated by the distance monitor described in Exer. 9, § 9.7. Very large-scale interferometers (Hough and Rowan 2000) are now being operated by a number of groups: LIGO in the USA (two 4-km detectors and a third that is 2 km long), VIRGO in Italy (3 km), GEO600 in Germany (600 m), and TAMA300 in Japan (300 m). These monitor the changes in separation between two pairs of heavy masses, suspended from supports that isolate the masses from outside vibrations. This approach has produced the most sensitive detectors to date, and is likely to produce the first detections. A very sensitive dedicated array of spacecraft, called LISA, is also planned. We will have much more to say about these detectors below.

Other ways of detecting gravitational waves are also being pursued.

- *Spacecraft tracking.* This principle has been used to search for gravitational waves using the communication data between Earth and interplanetary spacecraft (Armstrong 2006). By comparing small fluctuations in the round-trip time of radio signals sent to spacecraft, we try to identify gravitational waves. The sensitivity of these searches is not very high, however, because they are limited by the stability of the atomic clocks that are used for the timing and by delays caused by the plasma in the solar wind.

- *Pulsar timing.* Radio astronomers search for small irregularities in the times of arrival of signals from pulsars. Pulsars are spinning neutron stars that emit strong directed beams of radio waves, apparently because they have ultra-strong magnetic fields that are not aligned with the axis of rotation. Each time a magnetic pole happens to point toward Earth, the beamed emission is observed as a 'pulse' of radio waves. As we shall see in the next chapter, neutron stars can rotate very rapidly, even hundreds of times per second. Because the pulses are tied to the rotation rate, many pulsars are intrinsically very good clocks, potentially better than man-made ones (Cordes *et al.* 2004), and may accordingly be used for gravitational wave detection. Special-purpose pulsar timing arrays are currently searching for gravitational waves, by looking for correlated timing irregularities that could be caused by gravitational waves passing the radio array. When the planned Square Kilometer Array (SKA) radio telescope facility is built (perhaps around 2020), radio astronomers will have a superb tool for monitoring thousands of pulsars and digging deep for gravitational wave signals. But it is certainly conceivable that timing arrays may detect gravitational waves before the ground-based interferometers do.
- *Cosmic microwave background temperature perturbations.* Cosmologists study the fluctuations in the cosmic microwave background temperature distribution on the sky (see Ch. 12) for telltale signatures of gravitational waves from the Big Bang. The effect is difficult to measure, but it may come within reach of the Planck spacecraft, due for launch in 2009.

Interferometers, spacecraft tracking, and pulsar timing all share a common principle: they monitor electromagnetic radiation to look for the effects of gravitational waves. These are all examples of the general class of *beam detectors*.

It is important to bear in mind that these different approaches are suitable for different parts of the gravitational wave spectrum. Just as for the electromagnetic spectrum, gravitational waves at different frequencies carry different kinds of information. While ground-based detectors (bars and interferometers) typically observe between 10 Hz and a few kHz (although some prototypes have been built for MHz frequencies), space-based interferometers will explore the mHz part of the spectrum. Pulsar timing is suitable only at nHz frequencies, because we have to average over short-time fluctuations in the arrival times of pulses.

And the cosmic microwave background was affected by gravitational waves that had extremely long wavelengths at the time the universe was only a few hundred thousand years old (Ch. 12), so the frequencies of those waves today are of order 10^{-16} Hz!

A resonant detector

Resonant detectors are a good case study for the interaction of gravitational waves with continuous matter. To understand how they work in principle, we consider the following idealized detector, depicted in Fig. 9.2. Two point particles, each of mass m, are connected by a massless spring with spring constant k, damping constant ν, and unstretched length l_0. The system lies on the x axis of our TT coordinate system, with the masses at coordinate positions x_1 and x_2. In flat spacetime, the masses would obey the equations

$$mx_{1,00} = -k(x_1 - x_2 + l_0) - \nu(x_1 - x_2)_{,0} \tag{9.33}$$

and

$$mx_{2,00} = -k(x_2 - x_1 - l_0) - \nu(x_2 - x_1)_{,0}. \tag{9.34}$$

If we define the stretch ξ, resonant frequency ω_0, and damping rate γ by

$$\xi = x_2 - x_1 - l_0, \qquad \omega_0^2 = 2k/m, \qquad \gamma = \nu/m, \tag{9.35}$$

then we can combine Eqs. (9.33) and (9.34) to give

$$\xi_{,00} + 2\gamma\xi_{,0} + \omega_0^2\xi = 0, \tag{9.36}$$

the usual damped harmonic-oscillator equation.

What is the situation as a gravitational wave passes? We shall analyze the problem in three steps:

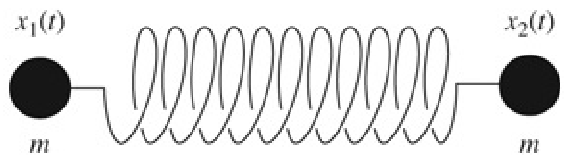


Figure 9.2 A spring with two identical masses as a detector of gravitational waves.

(1) A *free* particle remains at rest in the TT coordinates. This means that a local inertial frame at rest at, say, x_1, before the wave arrives remains at rest there after the wave hits. Let its coordinates be $\{x^{\alpha'}\}$. Suppose that the only motions in the system are those produced by the wave, *i.e.* that $\xi = 0(l_0|h_{\mu\nu}|) \ll l_0$. Then the masses' velocities will be small as well, and Newton's equations for the masses will apply in the local inertial frame:

$$mx^{j'}_{,0'0'} = F^{j'}, \tag{9.37}$$

where $\{F^{j'}\}$ are the components of any nongravitational forces on the masses. Because $\{x^{\alpha'}\}$ can differ from our TT coordinates $\{x^{\alpha}\}$ only by terms of order $h_{\mu\nu}$, and because x_1, $x_{1,0}$, and $x_{1,00}$ are all of order $h_{\mu\nu}$, we can use the TT coordinates in Eq. (9.37) with negligible error:

$$mx^{j}_{,00} = F^{j} + 0(|h_{\mu\nu}|^2). \tag{9.38}$$

(2) The only nongravitational force on each mass is that due to the spring. Since all the motions are slow, the spring will exert a force proportional to its instantaneous *proper* extension, as measured using the metric. If the proper

length of the spring is l, and if the gravitational wave travels in the z direction, then

$$l(t) = \int_{x_1(t)}^{x_2(t)} \left[1 + h_{xx}^{TT}(t)\right]^{1/2} \mathrm{d}x = [x_2(t) - x_1(t)]\left[1 + \tfrac{1}{2}h_{xx}^{TT}(t)\right] + 0(|h_{\mu\nu}|^2), \tag{9.39}$$

and Eq. (9.38) for our system gives

$$mx_{1,00} = -k(l_0 - l) - \nu(l_0 - l)_{,0}, \tag{9.40}$$

$$mx_{2,00} = -k(l - l_0) - \nu(l - l_0)_{,0}. \tag{9.41}$$

(3) Let us define the physical stretch ξ by

$$\xi = l - l_0. \tag{9.42}$$

We substitute Eq. (9.39) into this:

$$\xi = x_2 - x_1 - l_0 + \tfrac{1}{2}(x_2 - x_1)h_{xx}^{TT} + 0(|h_{\mu\nu}|^2). \tag{9.43}$$

Noting that the factor $(x_2 - x_1)$ multiplying h_{xx}^{TT} can be replaced by l_0 without changing the equation to the required order of accuracy, we can solve this to give

$$x_2 - x_1 = l_0 + \xi - \tfrac{1}{2}h_{xx}^{TT} l_0 + 0(|h_{\mu\nu}|^2). \tag{9.44}$$

If we use this in the difference between Eqs. (9.41) and (9.40), we obtain

$$\xi_{,00} + 2\gamma\xi_{,0} + \omega_0^2\xi = \tfrac{1}{2}l_0 h_{xx,00}^{TT}, \tag{9.45}$$

correct to first order in h_{xx}^{TT}. This is the fundamental equation governing the response of the detector to the gravitational wave. It has the simple form of a forced, damped harmonic oscillator. The forcing term is the tidal acceleration produced by the gravitational wave, as given in Eq. (9.28a), although our derivation started with the proper length computation in Eq. (9.24). This shows again the self-consistency of the two approaches to understanding the action of a gravitational wave on matter. An alternative derivation of this result using the equation of geodesic deviation may be found in Exer. 21, § 9.7. The generalization to waves incident from other directions is dealt with in Exer. 22, § 9.7.

We might use a detector of this sort as a resonant detector for sources of gravitational radiation of a fixed frequency (e.g. pulsars or close binary stars). (It can also be used to detect bursts – short wave packets of broad-spectrum radiation – but we will not discuss detecting those.) Suppose that the incident wave has the form

$$h_{xx}^{\mathrm{TT}} = A \cos \Omega t. \tag{9.46}$$

Then the steady solution of Eq. (9.45) for ξ is

$$\xi = R \cos(\Omega t + \phi), \tag{9.47}$$

with

$$R = \tfrac{1}{2} l_0 \Omega^2 A / [(\omega_0^2 - \Omega^2)^2 + 4\Omega^2\gamma^2]^{1/2}, \tag{9.48}$$

$$\tan\phi = 2\gamma\Omega/(\omega_0^2 - \Omega^2). \tag{9.49}$$

(Of course, the general initial-value solution for ξ will also contain transients, which damp away on a timescale $1/\gamma$.) The energy of oscillation of the detector is, to lowest order in h^{TT}_{xx},

$$E = \tfrac{1}{2} m (x_{1,0})^2 + \tfrac{1}{2} m (x_{2,0})^2 + \tfrac{1}{2} k \xi^2. \tag{9.50}$$

For a detector that was at rest before the wave arrived, we have $x_{1,0} = -x_{2,0} = -\xi_{,0}/2$ (see Exer. 23, § 9.7), so that

$$E = \tfrac{1}{4} m [(\xi_{,0})^2 + \omega_0^2 \xi^2] \tag{9.51}$$

$$= \tfrac{1}{4} m R^2 [\Omega^2 \sin^2(\Omega t + \phi) + \omega_0^2 \cos^2(\Omega t + \phi)]. \tag{9.52}$$

The mean value of this is its average over one period, $2\pi/\Omega$:

$$\langle E \rangle = \tfrac{1}{8} m R^2 (\omega_0^2 + \Omega^2). \tag{9.53}$$

We shall always use angle brackets $\langle\ \rangle$ to denote time averages.

If we wish to detect a specific source whose frequency Ω is known, then we should adjust ω_0 to equal Ω for maximum response (resonance), as we see from Eq. (9.48). In this case the amplitude of the response will be

$$R_{\text{resonant}} = \tfrac{1}{4} l_0 A (\Omega/\gamma) \tag{9.54}$$

and the energy of vibration is

$$E_{\text{resonant}} = \tfrac{1}{64} m l_0^2 \Omega^2 A^2 (\Omega/\gamma)^2. \tag{9.55}$$

The ratio Ω/γ is related to what is usually called the quality factor Q of an oscillator, where $1/Q$ is defined as the average fraction of the energy of the undriven oscillator that it loses (to friction) in one radian of oscillation (see Exer. 25, § 9.7):

$$Q = \omega_0/2\gamma. \tag{9.56}$$

In the resonant case we have

$$E_{\text{resonant}} = \tfrac{1}{16} m l_0^2 \Omega^2 A^2 Q^2. \tag{9.57}$$

What numbers are realistic for laboratory detectors? Most such detectors are massive cylindrical bars in which the ‘spring’ is the elasticity of the bar when it is stretched along its axis. When waves hit the bar broadside, they excite its longitudinal modes of vibration. The first detectors, built by Joseph Weber of the University of Maryland in the 1960s, were aluminum bars of mass 1.4×10^3 kg, length $l_0 = 1.5$ m, resonant frequency $\omega_0 = 10^4\ s^{-1}$, and Q about 10^5. This means that a strong resonant gravitational wave of $A = 10^{-20}$ (see § 9.3 below) will excite the bar to an energy of the order of 10^{-20} J. The resonant amplitude given by Eq. (9.54) is only about 10^{-15} m, roughly the diameter of an atomic nucleus! Many realistic gravitational waves will have amplitudes many orders of magnitude smaller than this, and will last for much too short a time to bring the bar to its full resonant amplitude.

Bar detectors in operation

When trying to measure such tiny effects, there are in general two ways to improve things: one is to increase the size of the effect, the other is to reduce any extraneous disturbances that might obscure the measurement. And then we have to determine how best to make the measurement. The size of the effect is controlled by the amplitude of the wave, the length of the bar, and the *Q*-value of the material. We can’t control the wave’s amplitude, and unfortunately extending the length is not an option: realistic bars may be as long as 3 m, but longer bars would be much harder to isolate from external disturbances. In order to achieve high values of *Q*, some bars have actually been made of single crystals, but it is hard to do better than that. Novel designs, such as spherical detectors that respond efficiently to waves from any direction, can increase the signal somewhat, but the difficulty of making bars intrinsically more sensitive is probably the main reason that they are not the detector of choice at the moment.

The other issue in detection is to reduce the extraneous noise. For example, thermal noise in any oscillator induces random vibrations with a mean energy of kT, where T is the absolute temperature and k is Boltzmann’s constant, $k = 1.38 \times 10^{-23}$ J/K.

In our example, this will be comparable to the energy of excitation if T is room temperature (∼300 K). But we chose a very optimistic wave amplitude. To detect reliably a wave with an amplitude ten times smaller would require a temperature 100 times smaller. For this reason, bar detectors in the 1980s began to change from room-temperature operation to cryogenic operation at around 3K. The coldest, and most sensitive, bar operating today is the Auriga bar, which goes below 100 mK.

Other sources of noise, such as vibrations from passing vehicles and everyday seismic disturbances, could be considerably larger than thermal noise, so the bar has to be very carefully isolated. This is done by hanging it from a support so that it forms a pendulum with a low resonant frequency, say 1 Hz. Vibrations from the ground may move the top attachment point of the pendulum, but little of this is transmitted through to the bar at frequencies above the pendulum frequency: pendulums are good low-pass mechanical filters. In practice, several sequential pendulums may be used, and the hanging frame is further isolated from vibration by using absorbing mounts.

How do resonant detectors measure such small disturbances? The measuring apparatus is called the transducer. Weber's original aluminum bar was instrumented with strain detectors around its waist, where the stretching of the metal is maximum. Other groups have tried to extract the energy of vibration from the bar into a transducer of very small mass that was resonant at the same frequency; if the energy extraction was efficient, then the transducer's amplitude of oscillation would be much larger. The most sensitive readout schemes involve ultra-low-noise low-temperature superconducting devices called SQUIDs.

We have confined our discussion to on-resonance detection of a continuous wave, in the case when there are no motions in the detector. If the wave comes in as a burst with a wide range of frequencies, where the excitation amplitude might be smaller than the broadband noise level, then we have to do a more careful analysis of their sensitivity, but the general picture does not change. One difficulty bars encounter with broadband signals is that it is difficult for them in practice (although not impossible in principle) to measure the frequency components of a waveform very far from their resonant frequencies, which normally lie above 600 Hz. Since most strong sources of gravitational waves emit at lower frequencies, this is a serious problem. A second difficulty is that, to reach a sensitivity to bursts of amplitude around 10^{-21} (which is the level that interferometers reached in 2005), bars need to conquer the so-called quantum limit. At these small excitations, the energy put into the vibrations of the bar by the wave is below one quantum (one phonon) of excitation of the resonant mode being used to detect them. The theory of how to detect below the quantum limit – of how to manipulate the Heisenberg uncertainty relation in a macroscopic object like a bar – is fascinating. But the challenge has not yet been met in practice, and is therefore another serious problem that bars face. For more details on all of these issues, see Misner *et al.* (1973), Smarr (1979), or Blair (1991).

The severe technical challenges of bar detectors come fundamentally from their small size: any detector based on the resonances of a metal object cannot be larger than a few meters in size, and that seriously limits the size of the tidal stretching induced by a gravitational wave. Laser interferometer detectors are built on kilometer scales (and in space, on scales of millions of kilometers). They therefore have an inherently larger response and are consequently able to go to a higher sensitivity before they become troubled by quantum, vibration, and thermal noise. The inherent difficulties faced by bars have led to a gradual reduction in research funding for bar detectors during the period after 2000, as interferometers have steadily improved and finally surpassed the sensitivity of the best bars. After 2010 it seems unlikely that any bar detectors will remain in operation.

Measuring distances with light

One of the most convenient ways of measuring the range to a distant object is by radar: send out a pulse of electromagnetic radiation, measure how long it takes to return after reflecting from the distant object, divide that by two and multiply by c, and that is the distance. Remarkably, because light occupies such a privileged position in the theory of relativity, this method is also an excellent way of measuring proper distances even in curved spacetime. It is the foundation of laser interferometric gravitational wave detectors.

We shall compute how to use light to measure the distance between two freely falling objects. Because the objects are freely falling, and because we make no assumption that they are close to one another, we shall use the TT coordinate system. Let us consider for simplicity at first a wave traveling in the z-direction with pure +–polarization, so that the metric is given by

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + [1 + h_+(z-t)]\mathrm{d}x^2 + [1 - h_+(t-z)]\mathrm{d}y^2 + \mathrm{d}z^2. \tag{9.58}$$

Suppose, again for simplicity, that the two objects lie on the x-axis, one of them at the origin $x = 0$ and the other at coordinate location $x = L$. In TT coordinates, they remain at these coordinate locations all the time. To make our measurement, the object at the origin sends a photon along the x-axis toward the other object, which reflects it back. The first object measures the amount of proper time that has elapsed since first emitting the photon. How is this related to the distance between the objects and to the metric of the gravitational wave?

Note that a photon traveling along the x-axis moves along a null world line ($\mathrm{d}s^2 = 0$) with $\mathrm{d}y = \mathrm{d}z = 0$. That means that it has an effective speed

$$\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2 = \frac{1}{1+h_+}. \tag{9.59}$$

Although this is not equal to one, this is just a *coordinate speed*, so it does not contradict relativity. A photon emitted at time t_{start} from the origin reaches any coordinate location x in a time $t(x)$; this is essentially what we are trying to solve for. The photon reaches the other end, at the fixed coordinate position $x = L$, at the coordinate time given by integrating the effective speed of light from Eq. (9.59):

$$t_{\rm far} = t_{\rm start} + \int_0^L \left[1 + h_+(t(x))\right]^{1/2} \mathrm{d}x. \tag{9.60}$$

This is an implicit equation since the function we want to find, $t(x)$, is inside the integral. But in linearized theory, we can solve this by using the fact that h_+ is small. Where $t(x)$ appears in the argument of h_+, we can use its flat-spacetime value, since corrections due to h_+ will only bring in terms of order h_+^2 in Eq. (9.60). So we set $t(x) = t_{\rm start} + x$ inside the integral and expand the square root. The result is the explicit integration $t_{far} = t_{\rm start} + L + \frac{1}{2}\int_0^L h_+(t_{\rm start} + x)\mathrm{d}x.$

The light is reflected back, and a similar argument gives the total time for the return trip:

$$t_{\rm return} = t_{\rm start} + 2L + \frac{1}{2}\int_0^L h_+(t_{\rm start} + x)\mathrm{d}x + \frac{1}{2}\int_0^L h_+(t_{\rm start} + L + x)\mathrm{d}x. \tag{9.61}$$

Note that coordinate time t in the TT coordinates is proper time, so that this equation gives a value that can be measured.

What does this equation tell us? First, suppose that L is actually rather small compared to the gravitational wavelength, or equivalently that the return time is small compared to the period of the wave, so that h_+ is effectively constant during the flight of the photon. Then the return time is just proportional to the proper distance to L as measured by this metric. This should not be surprising: for a small separation we could set up a local inertial frame in free fall with the particles, and in this frame all experiments should come out as they do in special relativity. In SR we know that radar ranging gives the correct proper distance, so it must do so here as well.

More generally, how do we use this equation to measure the metric of the wave? The simplest way to use it is to differentiate $t_{\rm return}$ with respect to $t_{\rm start}$, *i.e.* to monitor the rate of change of the return time as the wave passes. Since the only way that $t_{\rm start}$ enters the integrals in Eq. (9.61) is as the argument $t_{\rm start} + x$ of h_+, a derivative of the integral operates only as a derivative of h_+ with respect to its argument. Then the integration with respect to x is an integral of the derivative of h_+ over its argument, which simply produces h_+ again. The result of this is the vastly simpler expression

$$\frac{dt_{\text{return}}}{dt_{\text{start}}} = 1 + \tfrac{1}{2}\left[h_+(t_{\text{start}} + 2L) - h_+(t_{\text{start}})\right]. \tag{9.62}$$

This is rather a remarkable result: the rate of change of the return time depends only on the metric of the wave at the time the wave was emitted and when it was received back at the origin. In particular, the wave amplitude when the photon reflected off the distant object plays no role.

Now, if the signal sent out from the origin is not a single photon but a continuous electromagnetic wave with some frequency v, then each ‘crest’ of the wave can be thought of as another null ray or another photon being sent out and reflected back. The derivative of the time it takes these rays to return is nothing more than the change in the frequency of the electromagnetic wave:

$$\frac{dt_{\text{return}}}{dt_{\text{start}}} = \frac{\nu_{\text{return}}}{\nu_{\text{start}}}.$$

So if we monitor changes in the redshift of the returning wave, we can relate that directly to changes in the amplitude of the gravitational waves.

So far we have used a rather special arrangement of objects and wave: the wave was traveling in a direction perpendicular to the separation of the objects. If instead the wave is traveling at an angle θ to the z axis in the $x - -z$ plane, the return time derivative does involve the wave amplitude at the reflection time:

$$\begin{aligned}\frac{dt_{return}}{dt_{\text{start}}} = 1 + \tfrac{1}{2}\{&(1-\sin\theta)h_+(t_{\text{start}} + 2L) - (1+\sin\theta)h_+(t_{\text{start}}) \\ &+2\sin\theta h_+[t_{\text{start}} + (1-\sin\theta)L]\}.\end{aligned} \tag{9.63}$$

This three-term relation is the starting point for analyzing the response of beam detectors, as we shall see next. For its derivation see, Exer. 27, § 9.7.

Beam detectors

The simplest beam detector is spacecraft tracking (Armstrong 2006). Interplanetary spacecraft carry transponders, which are radio receivers that amplify and return the signals they receive from the ground tracking station. A measurement of the return time of the signal tells the space agency how far away the spacecraft is. If the measurement is accurate enough, small changes in the return time might be caused by gravitational waves. In practice, this is a difficult measurement to make, because fluctuations might also be caused by changes in the index of refraction of the thin interplanetary plasma or of the ionosphere, both of which the signals must pass through. These effects can be discriminated from a true gravitational wave by using the three-term relation, Eq. (9.63). The detected waveform has to appear in three different places in the data, once with the opposite sign. Random fluctuations are unlikely to do this.

Plasma fluctuations can also be suppressed by using multiple transponding frequencies (allowing them to be measured) or by using higher-frequency transponding, even using infrared laser light, which is hardly affected by plasma at all. But even then there will be another limit on the accuracy: the stability of the clock at the tracing station that measures the elapsed time t_{return}. Even the best clocks are, at present, not stable at the 10^{-19} level. It follows that a beam detector of this type could not expect to measure gravitational waves with amplitudes of 10^{-20} or below. Unfortunately, as we shall see below, this is where we expect almost all amplitudes to lie.

The remedy is to use an interferometer. In an interferometer, light from a stable laser passes through a beam splitter, which sends half the light down one arm and the other half down a perpendicular arm. The two beams of light have correlated phases. When they return after reflecting off mirrors at the ends of the arms, they are brought back into interference (see Fig. 9.3). The interference measures the difference in the armlengths of the two arms. If this difference changes, say because a gravitational wave passes through, then the interference pattern changes and the wave can in principle be detected.

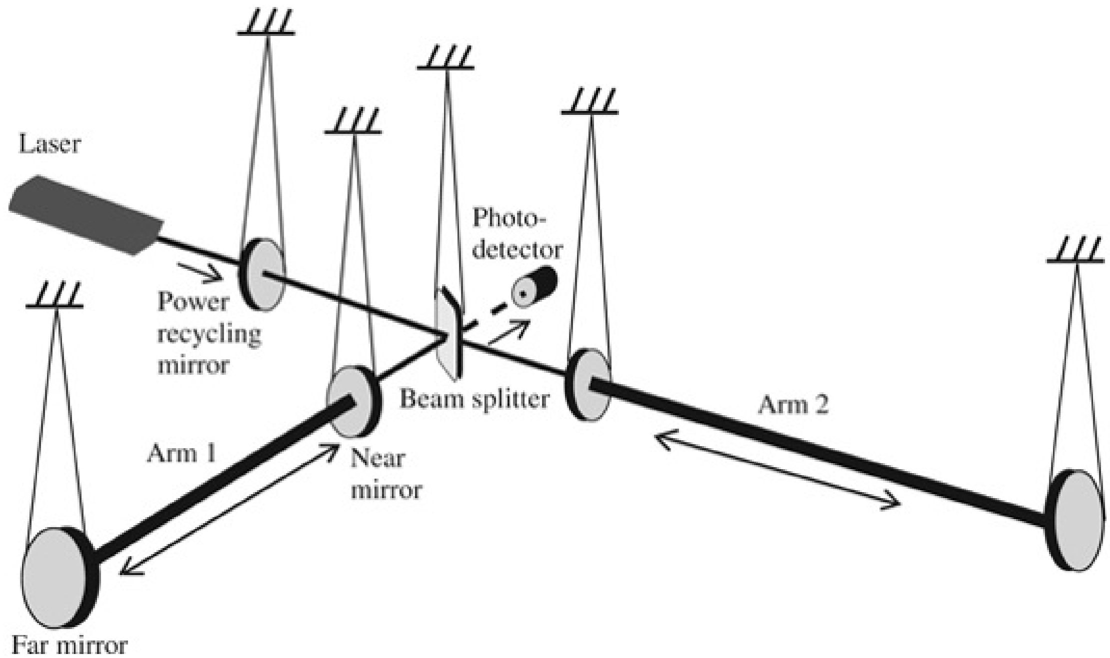


Figure 9.3 Sketch of the configuration of an interferometer like LIGO or VIRGO. Light from the laser passes through the power recycling mirror to the beamsplitter, where it is divided between the two arms. The arms form cavities, trapping most of the light, because the near mirrors are almost fully reflecting. This power buildup increases the sensitivity. The light exiting the cavities returns to the beamsplitter. A beam with destructive interference (difference of amplitudes) goes toward the photodetector; it should be dark unless a gravitational wave is present. The other return beam from the beamsplitter is the constructive interference beam returning toward the laser. Almost all the light goes here, and in order not to waste it, the power recycling mirror returns it to the interferometer in phase with the new incoming laser light. All the mirrors, including the beamsplitter, are suspended in order to filter out mechanical vibration noise. Other refinements of the optical design, such as mode cleaners and signal recycling mirrors, are not illustrated. The diagram is not to scale: the arms are 3 to 4 km long, while the central area (near mirrors, beamsplitter, laser, photodetector) is contained in a single building.

Now, an interferometer can usefully be thought of as two beam detectors laid perpendicular to one another. The two beams are correlated in phase: any given ‘crest’ of the light wave starts out down both arms at the same time. If the arms have the same proper length, then their beams will return in phase, interfering constructively. If the arms differ, say, by half the wavelength of light, then they will return out of phase and they will destructively interfere. An interferometer solves the problem that a beam detector needs a stable clock. The ‘clock’ in this case is one of the arms. The time-travel of the light in one arm essentially provides a reference for the return time of the light in the other arm. The technology of lasers, mirrors, and other systems is such that this reference can be used to measure changes in the return travel time of the other arm much more stably than the best atomic clocks would be able to do.

Interferometers are well-suited to registering gravitational waves: to see how this works, look at the polarization diagram Fig. (9.1). Imagine putting an interferometer in the circle in panel (a), with the beamsplitter at the center and the ends of the arms where the circle intersects the *x*- and *y*-directions. Then when a wave with +–polarization arrives, as in panel (b), it will stretch one arm and at the same time compress the other. This doubles the effective armlength difference that the interference pattern measures. But even if the wave arrives from, say, the *x*-direction, then its transverse action will still compress and expand the *y*-arm, giving a signal half as large as the maximum. But a wave of ×-polarization, as in panel (c), arriving perpendicular to both arms, will not stretch the interferometer arms at all, and so will not be detected. The interferometer is therefore a linearly polarized detector that responds to signals arriving from almost all directions.

Because interferometers bring light beams together and interfere them, it is often mistakenly thought that they measure the changes in their armlengths against a standard wavelength, and that in turn sometimes leads students to ask whether the wavelength of light is affected by the gravitational wave, thereby invalidating the measurement. Our discussion of beam detectors should make clear that the wavelength of the light is not the relevant thing: interferometers basically compare two *return times*, and as long as light travels up and down the arms on a null worldline, it does not matter at all what the wavelength is doing along the way.[1]

Interferometer observations

Interferometers are now (2008) the most sensitive operating detectors, and they are the most likely instruments to make the first direct detections. They have two major advantages over bar detectors: their sensitivity can be increased considerably without running into fundamental problems of materials or physics, and (because they do not depend on any resonant vibration) they operate over a broad spectrum of frequencies.

In designing an interferometer for gravitational waves, scientists face the same basic options as for bar detectors: try to increase the size of the signal and try to reduce the extraneous noise. Unlike bars, interferometers have a natural way of increasing the signal: extending their length. The longer the arm, the larger will be the difference in return times for a given gravitational wave amplitude. The largest earth-bound detectors are the two 4-km LIGO detectors in the USA, closely followed by 3-km VIRGO in Italy. But physical length is only the first step. The light beams inside the arms can be folded over multiple passes up and down or contained within resonating light cavities, so that the residence time of light in each arm is longer; this again increases the difference in return times when there is a gravitational wave. A simplified sketch of the optical design of the LIGO or VIRGO detector is shown in Fig. (9.3).

But even with long arms, the signal can be masked by a variety of instrumental sources of noise (Saulson 1994, Hough and Rowan 2000). To filter out vibrations from ground disturbances, interferometers use the same strategy as bar detectors: the optical components are suspended. Two-or three-stage suspensions are normal, but VIRGO uses seven pendulums, each hanging from the one above; this is designed to allow the instrument to observe at lower frequencies, where seismic vibration noise is stronger. Thermal vibrations of mirrors and their suspensions are, as for bars, a serious problem, but interferometers do not have the option to operate at temperatures as low as 4K, because the heat input from the laser beam on the mirrors would be impossible to remove. Current interferometers (as of 2008) all operate at room temperature, although there are plans in Japan to pioneer operation at 40K. (The project is known as LCGT.) Thermal noise is controlled by using ultra-high-Q materials for the mirrors and suspension wires; this confines the thermal noise to narrow bands around the resonant frequencies of the mirror vibrations and pendulum modes, and these are designed to be well outside the observation band of the instruments. This approach, which includes using optical fiber as the suspension wire and bonding it to mirrors without glue, has been pioneered in the GEO600 detector. As interferometers become more advanced, they will also have to contend with quantum sources of noise and the Heisenberg uncertainty principle on the mirror locations, but even here there is a clearer way of solving these problems of the quantum limit than for bar detectors.

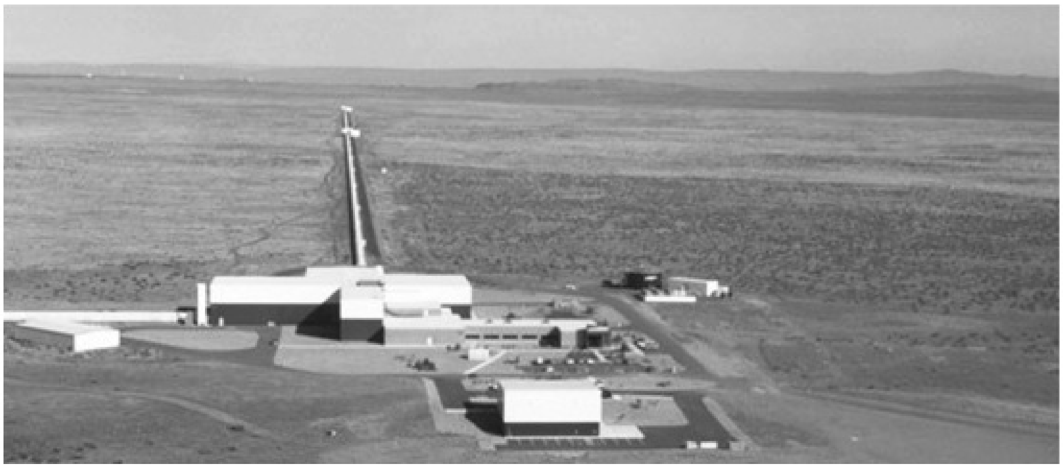


Figure 9.4 The LIGO gravitational wave observatory at Hanford, Washington, DC. One of the 4-km arms stretches into the distance, the other leaves the photo off to the left. The laser light in the arms is contained inside 4-km-long stainless-steel vacuum tubes 1.6 m in diameter. This observatory actually operates two interferometers, the second one only 2 km long. The building housing its end mirror can be seen half-way along the arm. The shorter instrument helps discriminate against local disturbances. For an aerial view of this site go to a geographic display engine, such as Google Maps, and type in the longitude and latitude (46.45, −119.4). The other LIGO detector (which has only one interferometer) is in Livingston, Louisiana, at (30.55, −90.79). The VIRGO detector near Pisa is at (43.64, 10.50), and GEO600 detector near Hanover is at (52.25, 9.81). (Photo courtesy LIGO Laboratory.)

The final source of noise is what physicists call shot noise, which is the random fluctuations of intensity in the interference of the two beams that comes from the fact that the beams are composed of discrete photons and not continuous classical electromagnetic radiation. Shot noise is the major limitation on sensitivity in interferometers at frequencies above about 200 Hz. It can be reduced, and hence the sensitivity increased, by increasing the amount of light stored in the arm cavities, because with more photons the power fluctuations go down.

The envelope of the different noise sources provides an observing band for ground-based detectors that, for current instruments, runs from about 40 Hz up to around 1 kHz. At low frequencies, it is difficult for the suspensions to isolate the mirrors from ground vibration. In the middle of this range, the sensitivity limit is set by thermal noise from suspensions and mirror vibrations. At higher frequencies, the limit is shot noise.

The detection of gravitational waves involves more than building and operating sensitive detectors. It also requires appropriate data analysis, because the expected signals will be below the broadband noise and must be extracted by intelligent computer-based filtering of the data. Because the detectors are so complex, there is always the possibility that random internal disturbances will masquerade as gravitational wave signals, so in practice signals need to be seen in more than one detector at the same time in order for the scientists to have confidence. Then the tiny (millisecond-scale) time-delays between different detectors' observations provide information on the location of the source on the sky.

Between mid-2005 and late 2007, the LIGO detectors logged operation in coincidence (all three detectors) for more than a year at a sensitivity better than 10^{-21} for broadband bursts of gravitational waves. From what astronomers know about potential sources of gravitational waves (see below), it is certainly possible that signals of this strength would arrive once or twice per year, but it is also possible that they are as rare as once per hundred years! GEO600 has operated for more than half of this same period in coincidence as well. VIRGO began operation at a similar sensitivity in early 2007. The LIGO and GEO detectors pool their data and analyze them jointly in an organization called the LIGO Scientific Collaboration (LSC). VIRGO also shares its data, which are then analyzed jointly with LSC data. If further large-scale detectors are brought into operation (there are advanced plans in Japan, as mentioned earlier, and in Australia), then they will presumably also join these efforts. Each new detector improves the sensitivity of all existing detectors.

The existing detector groups plan modest incremental upgrades in sensitivity during the remainder of the first decade of the twenty-first century, and then LIGO and VIRGO expect to upgrade to sensitivities better than 10^{-22} and to push their lower frequency limit closer to 10 Hz. These major upgrades, called Advanced LIGO and Advanced VIRGO, will involve many new components and much more powerful lasers. As we will see below, regular detections of gravitational waves are almost guaranteed at that point. But the first detection could of course come at any time during this development schedule.

Even more ambitious than the ground-based detector projects is the LISA mission, a joint undertaking of the European Space Agency (ESA) and the US space agency NASA that is currently planned for launch around 2018. Going into space is necessary if we want to observe at frequencies below about 1 Hz. At these low frequencies, the Earth's Newtonian gravitational field is too noisy: any change in gravity will be registered by detectors, and even the tiny changes in gravity associated with the density changes of seismic waves and weather systems are larger than the expected amplitudes of gravitational waves. So low-frequency observing needs to be done far from the Earth.

LISA will consist of three spacecraft in an equilateral triangle, all orbiting the Sun at a distance of 1 AU, the same as the Earth, and trailing the Earth by 20°. Their separation will be 5×10^6 km, well-matched to detecting gravitational waves in the millihertz frequency range. The three arms can be combined in various ways to form three different two-armed interferometers, which allows LISA to measure both polarizations of an incoming wave and to sweep the sky with a fairly uniform antenna pattern. As with ground-based instruments, LISA must contend with noise. Thermal noise is not an issue because its large armlength means that the signal it is measuring – the time-difference between arms – is much larger than would be induced by vibrations of materials. But external disturbances, caused by the Sun's radiation pressure and the solar wind, are significant, and so the LISA spacecraft must be designed to fall freely to high accuracy. Each spacecraft contains two free masses (called proof masses) that are undisturbed and able to follow geodesics. The spacecraft senses the positions of the masses and uses very weak jets to adjust its position so it does not disturb the proof masses. The proof masses are used as the reference points for the interferometer arms. This technique is called drag-free operation, and is one of a number of fascinating technologies that LISA will pioneer.

Communicating with 1 W lasers, LISA achieves a remarkable sensitivity, and will be able to see the strongest sources in its band anywhere in the universe. We will return to the kinds of sources that might be detected by LISA and the ground-based detectors after studying the way that gravitational waves are generated in general relativity.

9.3 The generation of gravitational waves

Simple estimates

It is easy to see that the amplitude of any gravitational waves incident on Earth should be small. A ‘strong’ gravitational wave would have $h_{\mu\nu} = 0(1)$, and we should expect amplitudes like this only very near a highly relativistic source, where the Newtonian potential (if it had any meaning) would be of order one. For a source of mass *M*, this would be at distances of order *M* from it. As with all radiation fields, the amplitude of the gravitational waves falls off as r^{-1} far from the source. (Readers who are not familiar with solutions of the wave equation will find demonstrations of this in the next sections.) So if Earth is a distance *R* from a source of mass *M*, the largest amplitude waves we should expect are of order *M*/*R*. For the formation of a 10 $M_{\odot}$ black hole in a supernova explosion in a nearby galaxy 10^{23} m away, this is about 10^{-17}. This is in fact an upper limit in this case, and less-violent events will lead to very much smaller amplitudes.

Slow motion wave generation

Our object is to solve Eq. (8.42):

$$\left(-\frac{\partial^2}{\partial t^2} + \nabla^2\right)\bar{h}_{\mu\nu} = -16\pi T_{\mu\nu}. \tag{9.64}$$

We will find the exact solution in a later section. Here we will make some simplifying – but realistic – assumptions. We assume that the time-dependent part of $T_{\mu\nu}$ is in sinusoidal oscillation with frequency Ω, *i.e.* that it is the real part of

$$T_{\mu\nu} = S_{\mu\nu}(x^i)\,\mathrm{e}^{-\mathrm{i}\Omega t}, \tag{9.65}$$

and that the region of space in which $S_{\mu\nu} \neq 0$ is small compared with $2\pi/\Omega$, the wavelength of a gravitational wave of frequency Ω. The first assumption is not much of a restriction, since a general time dependence can be reduced to a sum over sinusoidal motions by Fourier analysis. Besides, many interesting astrophysical sources *are* roughly periodic: pulsating stars, pulsars, binary systems. The second assumption is called the slow-motion assumption, since it implies that the typical velocity inside the source region, which is Ω times the size of that region, should be much less than one. All but the most powerful sources of gravitational waves probably satisfy this condition.

Let us look for a solution for $\bar{h}_{\mu\nu}$ of the form

$$\bar{h}_{\mu\nu} = B_{\mu\nu}(x^i)\,\mathrm{e}^{-\mathrm{i}\Omega t}. \tag{9.66}$$

(In the end we must take the real part of this for our answer.) Putting this and Eq. (9.65) into Eq. (9.64) gives

$$(\nabla^2 + \Omega^2)B_{\mu\nu} = -16\pi S_{\mu\nu}. \tag{9.67}$$

It is important to bear in mind as we proceed that the indices on $\bar{h}_{\mu\nu}$ in Eq. (9.64) play almost no role. We shall regard each component $\bar{h}_{\mu\nu}$ as simply a function on Minkowski space, satisfying the wave equation. All our steps will be the same as for solving the scalar equation $(-\partial^2/\partial t^2 + \nabla^2)f = g$, until we come to Eq. (9.75).

Outside the source (i.e. where $S_{\mu\nu} = 0$) we want a solution $B_{\mu\nu}$ of Eq. (9.67) that represents outgoing radiation far away; and of all such possible solutions we want the one that dominates in the slow-motion limit. Let us define r to be the spherical polar radial coordinate, where the origin is chosen inside the source.

We show in Exer. 29, § 9.7 that the solution we seek is spherical outside the source,

$$B_{\mu\nu} = \frac{A_{\mu\nu}}{r}\,\mathrm{e}^{\mathrm{i}\Omega r} + \frac{Z_{\mu\nu}}{r}\,\mathrm{e}^{-\mathrm{i}\Omega r}, \tag{9.68}$$

where $A_{\mu\nu}$ and $Z_{\mu\nu}$ are constants. The term in $e^{-i\Omega r}$ represents a wave traveling toward the origin $r = 0$ (called an ingoing wave), while the other term is outgoing (see Exer. 28, § 9.7). We want waves emitted by the source, so we choose $Z_{\mu\nu} = 0$.

Our problem is to determine $A_{\mu\nu}$ in terms of the source. Here we make our approximation that the source is nonzero only inside a sphere of radius $\varepsilon \ll 2\pi/\Omega$. Let us integrate Eq. (9.67) over the interior of this sphere. One term we get is

$$\int \Omega^2 B_{\mu\nu}\,\mathrm{d}^3x \leqslant \Omega^2 |B_{\mu\nu}|_{\max} 4\pi\varepsilon^3/3, \tag{9.69}$$

where $|B_{\mu\nu}|_{\max}$ is the maximum value $B_{\mu\nu}$ takes inside the source. We will see that this term is negligible. The other term from integrating the left-hand side of Eq. (9.67) is

$$\int \nabla^2 B_{\mu\nu}\,\mathrm{d}^3x = \oint \boldsymbol{n}\cdot\nabla B_{\mu\nu}\,\mathrm{d}S, \tag{9.70}$$

by Gauss' theorem. But the surface integral is outside the source, where only the spherical part of $B_{\mu\nu}$ given by Eq. (9.68) survives:

$$\oint \boldsymbol{n}\cdot\nabla B_{\mu\nu}\,\mathrm{d}S = 4\pi\varepsilon^2\left(\frac{\mathrm{d}}{\mathrm{d}r}B_{\mu\nu}\right)_{r=\varepsilon} \approx -4\pi A_{\mu\nu}, \tag{9.71}$$

again with the approximation $\varepsilon \ll 2\pi/\Omega$. Finally, we define the integral of the right-hand side of Eq. (9.67) to be

$$J_{\mu\nu} = \int S_{\mu\nu}\,\mathrm{d}^3x. \tag{9.72}$$

Combining these results in the limit $\varepsilon \to 0$ gives

$$A_{\mu\nu} = 4J_{\mu\nu}, \tag{9.73}$$

$$\bar{h}_{\mu\nu} = 4J_{\mu\nu}\,\mathrm{e}^{\mathrm{i}\Omega(r-t)}/r. \tag{9.74}$$

These are the expressions for the gravitational waves generated by the source, neglecting terms of order r^{-2} *and* any r^{-1} terms that are higher order in $\varepsilon\Omega$.

It is possible to simplify these considerably. Here we begin to use the fact that $\{\bar{h}_{\mu\nu}\}$ are components of a single tensor, not the unrelated functions we have solved for in Eq. (9.74). From Eq. (9.72) we learn

$$J_{\mu\nu}\ \mathrm{e}^{-\mathrm{i}\Omega t} = \int T_{\mu\nu}\ \mathrm{d}^3x, \tag{9.75}$$

which has as one consequence:

$$-\,\mathrm{i}\Omega J^{\mu 0}\ \mathrm{e}^{-\mathrm{i}\Omega t} = \int T^{\mu 0}{}_{,0}\,\mathrm{d}^3x. \tag{9.76}$$

Now, the conservation law for $T_{\mu\nu}$,

$$T^{\mu\nu}{}_{,\nu} = 0, \tag{9.77}$$

implies that

$$T^{\mu 0}{}_{,0} = -T^{\mu k}{}_{,k} \tag{9.78}$$

and hence that

$$\mathrm{i}\Omega J^{\mu 0}\ \mathrm{e}^{-\mathrm{i}\Omega t} = \int T^{\mu k}{}_{,k}\ \mathrm{d}^3x = \oint T^{\mu k}\, n_k\ \mathrm{d}S, \tag{9.79}$$

the last step being the application of Gauss' theorem to any volume completely containing the source. This means that $T^{\mu\nu} = 0$ on the surface bounding this volume, so that the right-hand side of Eq. (9.79) vanishes. This means that if $\Omega \neq 0$, we have

$$J^{\mu 0} = 0, \qquad \bar{h}^{\mu 0} = 0. \tag{9.80}$$

These conditions basically embody the laws of conservation of total energy and momentum for the oscillating source. The neglected higher-order parts of $\bar{h}^{\mu 0}$ are gauge terms (Exer. 32, § 9.7).

The expression for J_{ij} can also be rewritten in an instructive way by using the result of Exer. 23, § 4.10.

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} \int T^{00} x^l x^m \, \mathrm{d}^3 x = 2 \int T^{lm} \, \mathrm{d}^3 x. \tag{9.81}$$

For a source in slow motion, we have seen in Ch. 7 that $T^{00} \approx \rho$, the Newtonian mass density. It follows that the integral on the left-hand side of Eq. (9.81) is what is often referred to as the quadrupole moment tensor of the mass distribution,

$$I^{lm} := \int T^{00} x^l x^m \, \mathrm{d}^3 x \tag{9.82a}$$

$$= D^{lm} \, \mathrm{e}^{-\mathrm{i}\Omega t} \tag{9.82b}$$

(Conventions for defining the quadrupole moment vary from one text to another. We follow Misner *et al.* (1973).) In terms of this we have

$$\bar{h}_{jk} = -2\Omega^2 D_{jk} \, \mathrm{e}^{\mathrm{i}\Omega(r-t)}/r. \tag{9.83}$$

It is important to remember that Eq. (9.83) is an approximation which neglects not merely all terms of order r^{-2} but also r^{-1} terms that are not dominant in the slow-motion approximation. In particular, $\bar{h}_{jk}{}^{,k}$ is of higher order, and this guarantees that the gauge condition $\bar{h}^{\mu\nu}{}_{,\nu} = 0$ is satisfied by Eqs. (9.83) and (9.80) at the lowest order in r^{-1} and Ω. Because of Eq. (9.83), this approximation is often called the *quadrupole* approximation for gravitational radiation.

As for the plane waves we studied earlier, we have here the freedom to make a further restriction of the gauge. The obvious choice is to try to find a TT gauge, transverse to the direction of motion of the wave (the radial direction), which has the unit vector $n^j = x^j/r$. Exer. 32, § 9.7, shows that this is possible, so that in the TT gauge we have the simplest form of the wave. If we choose our axes so that at the point where we measure the wave it is traveling in the z direction, then we can supplement Eq. (9.80) by

$$\bar{h}_{zi}^{\mathrm{TT}} = 0, \tag{9.84}$$

$$\bar{h}_{xx}^{\mathrm{TT}} = -\bar{h}_{yy}^{\mathrm{TT}} = -\Omega^2(\text{Ɨ}_{xx} - \text{Ɨ}_{yy})\, \mathrm{e}^{\mathrm{i}\Omega r}/r, \tag{9.85}$$

$$\bar{h}_{xy}^{\mathrm{TT}} = -2\Omega^2 \text{Ɨ}_{xy}\, \mathrm{e}^{\mathrm{i}\Omega r}/r, \tag{9.86}$$

where

$$\rlap{-}I_{jk} := I_{jk} - \tfrac{1}{3}\delta_{jk} I^{l}_{l} \tag{9.87}$$

is called the trace-free or reduced quadrupole moment tensor.

Examples

Let us consider the waves emitted by a simple oscillator like the one we used as a detector in § 9.2. If both masses oscillate with angular frequency ω and amplitude A about mean equilibrium positions a distance l_0, apart, then, by Exer. 30, § 9.7, the quadrupole tensor has only one nonzero component,

$$\begin{aligned} I_{xx} &= m[(x_1)^2 + (x_2)^2] \\ &= [(-\tfrac{1}{2}l_0 - A\cos\omega t)^2 + (\tfrac{1}{2}l_0 - A\cos\omega t)^2] \\ &= \text{const.} + mA^2 \cos 2\omega t + 2ml_0 A\cos\omega t. \end{aligned} \tag{9.88}$$

Recall that only the sinusoidal part of I_{xx} should be used in the formulae developed in the previous paragraph. In this case, there are two such pieces, with frequencies ω and 2ω. Since the wave equation Eq. (9.64) is linear, we shall treat each term separately and simply add the results later. The ω term in I_{xx} is the real part of $2ml_0A\exp(-i\omega t)$. The trace-free quadrupole tensor has components

$$\left.\begin{aligned} \bar{I}_{xx} &= I_{xx} - \tfrac{1}{3}I^j{}_j = \tfrac{2}{3}I_{xx} = \tfrac{4}{3}ml_0A\,\mathrm{e}^{-\mathrm{i}\omega t}, \\ \bar{I}_{yy} &= \bar{I}_{zz} = -\tfrac{1}{3}I_{xx} = -\tfrac{2}{3}ml_0A\,\mathrm{e}^{-\mathrm{i}\omega t}, \end{aligned}\right\} \tag{9.89}$$

all off-diagonal components vanishing. If we consider the radiation traveling in the z direction, we get, from Eqs. (9.84)–(9.86),

$$\bar{h}^{\mathrm{TT}}_{xx} = -\bar{h}^{\mathrm{TT}}_{yy} = -2m\omega^2 l_0 A\,\mathrm{e}^{\mathrm{i}\omega(r-t)}/r, \qquad \bar{h}^{\mathrm{TT}}_{xy} = 0. \tag{9.90}$$

The radiation is linearly polarized, with an orientation such that the ellipse in Fig. 9.1 is aligned with the line joining the two masses. The same is true for the radiation going in the y direction, by symmetry. But for the radiation traveling in the x direction (i.e. along the line joining the masses), we need to make the substitutions $z \to x$, $x \to y$, $y \to z$ in Eqs. (9.85)–(9.86), and we find

$$\bar{h}^{\mathrm{TT}}_{ij} = 0. \tag{9.91}$$

There is *no* radiation in the x direction. In Exer. 36, § 9.7, we will fill in this radiation pattern by calculating the amount of radiation and its polarization in arbitrary directions.

A similar calculation for the 2ω piece of I_{xx} gives the same radiation pattern, replacing Eq. (9.90) by

$$\bar{h}^{\mathrm{TT}}_{xx} = -\bar{h}^{\mathrm{TT}}_{yy} = -4\, m\omega^2 A^2\ \mathrm{e}^{2\mathrm{i}\omega(r-t)}/r, \qquad \bar{h}^{\mathrm{TT}}_{xy} = 0. \tag{9.92}$$

The total radiation field is the real part of the sum of Eqs. (9.90) and (9.92),

$$\bar{h}^{\mathrm{TT}}_{xx} = -[2\, m\, \omega^2 l_0 A \cos \omega(r-t) + 4\, m\, \omega^2 A^2 \cos 2\, \omega(r-t)]/r. \tag{9.93}$$

Let us estimate the radiation from a laboratory-sized generator of this type. If we take $m = 10^3$ kg $= 7 \times 10^{-24}$ m, $l_0 = 1$ m, $A = 10^{-4}$ m, and $\omega = 10^4\ \mathrm{s}^{-1} = 3 \times 10^{-4}\ \mathrm{m}^{-1}$, then the 2ω contribution is negligible and we find that the amplitude is about $10^{-34}/r$, where r is measured in meters. This shows that laboratory generators are unlikely to produce useful gravitational waves in the near future!

A more interesting example of a gravitational wave source is a binary star system. Strictly speaking, our derivation applies only to sources where motions result from nongravitational forces (this is the content of Eq. (9.77)), but our final result, Eqs. (9.84)–(9.87), makes use only of the motions produced, not of the forces. It is perhaps not so surprising, then, that Eqs. (9.84)–(9.87) are a good first approximation even for systems dominated by Newtonian gravitational forces (see further reading for references). Let us suppose, then, that we have two stars of mass m, idealized as points in circular orbit about one another, separated a distance l_0 (i.e. moving on a circle of radius $\frac{1}{2}l_0$). Their orbit equation (gravitational force = ‘centrifugal force’) is

$$\frac{m^2}{l_0^2} = m\omega^2 \left(\frac{l_0}{2}\right) \Rightarrow \omega = (2\, m/l_0^3)^{1/2}, \tag{9.94}$$

where ω is the angular velocity of the orbit. Then, with an appropriate choice of coordinates, the masses move on the curves

$$\left.\begin{aligned} x_1(t) &= \tfrac{1}{2} l_0 \cos \omega t, & y_1(t) &= \tfrac{1}{2} l_0 \sin \omega t, \\ x_2(t) &= -x_1(t), & y_2(t) &= -y_1(t), \end{aligned}\right\} \tag{9.95}$$

where the subscripts 1 and 2 refer to the respective stars. These equations give

$$\left.\begin{aligned} I_{xx} &= \tfrac{1}{4}ml_0^2 \cos 2\,\omega t + \text{const.},\\ I_{yy} &= -\tfrac{1}{4}ml_0^2 \cos 2\,\omega t + \text{const.},\\ I_{xy} &= \tfrac{1}{4}ml_0^2 \sin 2\,\omega t. \end{aligned}\right\} \tag{9.96}$$

The only nonvanishing components of the reduced quadrupole tensor are, in complex notation and omitting time-independent terms,

$$\left.\begin{aligned} \rlap{-}I_{xx} &= -\rlap{-}I_{yy} = \tfrac{1}{4}ml_0^2\ \mathrm{e}^{-2\mathrm{i}\omega t},\\ \rlap{-}I_{xy} &= \tfrac{1}{4}\mathrm{i}ml_0^2\ \mathrm{e}^{-2\mathrm{i}\omega t}. \end{aligned}\right\} \tag{9.97}$$

All the radiation comes out with frequency $\Omega = 2\,\omega$. The radiation along the z direction (perpendicular to the plane of the orbit) is, by Eqs. (9.84)–(9.86),

$$\left.\begin{aligned} \bar{h}_{xx} &= -\bar{h}_{yy} = -2\,ml_0^2\,\omega^2\ \mathrm{e}^{2\mathrm{i}\omega(r-t)}/r,\\ \bar{h}_{xy} &= -2\,\mathrm{i}\,ml_0^2\,\omega^2\ \mathrm{e}^{2\mathrm{i}\omega(r-t)}/r. \end{aligned}\right\} \tag{9.98}$$

This is *circularly polarized* radiation (see Exer. 15, § 9.7). The radiation in the plane of the orbit, say in the x direction, is found in the same manner used to derive Eq. (9.91). This gives

$$\bar{h}^{\mathrm{TT}}_{yy} = -\bar{h}^{\mathrm{TT}}_{zz} = ml_0^2\omega^2\ \mathrm{e}^{2\mathrm{i}\omega(r-t)}/r, \tag{9.99}$$

all others vanishing. This shows linear polarization aligned with the orbital plane. The antenna pattern and polarization are examined in greater detail in Exer. 38, § 9.7, and the calculation is generalized to unequal masses in elliptical orbits in Exer. 39.

The amplitude of the radiation is of order $ml_0^2\omega^2/r$, which, by Eq. (9.94), is of order $(m\omega)^{2/3}m/r$. Binary systems are likely to be very important sources of gravitational waves, indeed they may well be the first sources detected. Binary systems containing pulsars have already provided strong indirect evidence for gravitational radiation (Stairs 2003, Lorimer 2008). The first, and still the most important of these systems is the one containing the pulsar PSR B1913+16, the discovery of which by Hulse and Taylor (1975) led to their being awarded the Nobel Prize for Physics in 1993. This system consists of two neutron stars (see Ch. 10) orbiting each other closely. The orbital period, inferred from the Doppler shift of the pulsar's period, is 7 h 45 min 7 s (27907 s or 8.3721×10^{12} m), and both stars have masses approximately equal to $1.4M_{\odot}$ (2.07 km) (Taylor and Weisberg 1982). If the system is 8 kpc = 2.4×10^{20} m away, then its radiation will have the approximate amplitude 10^{-23} at Earth. We will calculate the effect of this radiation on the binary orbit itself later in this chapter. In Ch. 10 we will discuss the dynamics of the system, including how the masses are measured.

Order-of-magnitude estimates

Although our simple approach does not enable us to write down solutions for $\bar{h}_{\mu\nu}$ generated by more complicated, nonperiodic motions, we can use Eq. (9.83) to obtain some order-of-magnitude estimates. Since D_{jk} is of order MR^2, for a system of mass M and size R, the radiation will have amplitude about $M(\Omega R)^2/r \approx v^2(M/r)$, where v is a typical internal velocity in the source. This applies directly to Eq. (9.99); note that in Eq. (9.93) the first term uses, instead of R^2, the product $l_0 A$ of the two characteristic lengths in the problem. If we are dealing with, say, a collapsing mass moving under its own gravitational forces, then by the virial theorem $v^2 \sim \phi_0$, the typical Newtonian potential in the source, while $M/r \sim \phi_r$, the Newtonian potential of the source at the observer's distance r. Then we get the simple upper limit

$$h \lesssim \phi_0 \phi_r. \tag{9.100}$$

So the wave amplitude is always less than, or of the order of, the Newtonian potential ϕ_r. Why then can we detect h but not ϕ_r itself; why can we hope to find waves from a supernova in a distant galaxy without being able to detect its presence gravitationally before the explosion? The answer lies in the forces involved. The Newtonian tidal gravitational force on a detector of size l_0 at a distance r is about $\phi_r l_0/r^2$, while the wave force is $hl_0\omega^2$ (see Eq. (9.45)). The wave force is thus a factor $\phi_0(\omega r)^2 \sim (\phi_0 r/R)^2$ larger than the Newtonian force. For a relativistic system ($\phi_0 \sim 0.1$) of size 1 AU ($\sim 10^{11}$ m), observed by a detector a distance 10^{23} m away, this factor is 10^{22}. This estimate, incidentally, gives the largest distance r at which we may still approximate the gravitational field of a dynamical system as Newtonian (i.e. neglecting wave effects): $r = R/\phi_0$, where R is the size of the system.

The estimate in Eq. (9.100) is really an optimistic upper limit, because it assumed that all the mass motions contributed to D_{jk}. In realistic situations this could be a serious overestimate because of the following fundamental fact: *spherically symmetric motions do not radiate*. The rigorous proof of this is discussed in Ch. 10, but in Exer. 40, § 9.7 we derive it from linearized theory, Eq. (9.101) below. It also seems to follow from Eq. (9.82a): if T^{00} is spherically symmetric, then I^{lm} is proportional to δ^{lm} and Ɨ^{lm} vanishes. But this argument has to be treated with care, since Eq. (9.82a) is part of an approximation designed to give only the dominant radiation. We would have to show that spherically symmetric motions would not contribute to terms of higher order in the approximation if they were present. This is in fact true, and it is interesting to ask what eliminates them. The answer is Eq. (9.77): conservation of energy eliminates ‘monopole’ radiation in linearized theory, just as conservation of charge eliminates monopole radiation in electromagnetism.

The danger of using Eq. (9.82a) too glibly is illustrated in Exer. 31e, § 9.7: four equal masses at the corners of a rotating square give no time-dependent I^{lm} and hence no radiation in this approximation. But they *would* emit radiation at a higher order of approximation.

Exact solution of the wave equation

Readers who have studied the wave equation, Eq. (9.64), will know that its outgoing-wave solution for arbitrary $T_{\mu\nu}$ is given by the retarded integral

$$\begin{aligned} \bar{h}_{\mu\nu}(t, x^i) &= 4 \int \frac{T_{\mu\nu}(t - R, y^i)}{R}\, \mathrm{d}^3 y, \\ R &= |x^i - y^i|, \end{aligned} \tag{9.101}$$

where the integral is over the past light cone of the event (t, x^i) at which $\bar{h}_{\mu\nu}$ is evaluated. We let the origin be inside the source and we suppose that the field point x^i is far away,

$$|x^i| := r \gg |y^i| := y, \tag{9.102}$$

and that time derivatives of $T_{\mu\nu}$ are small. Then, inside the integral, Eq. (9.101), the dominant contribution comes from replacing R^{-1} by r^{-1}:

$$\bar{h}_{\mu\nu}(t, x^i) \approx \frac{4}{r} \int T_{\mu\nu}(t - R, y^i)\, \mathrm{d}^3 y. \tag{9.103}$$

This is the generalization of Eq. (9.74). Now, by virtue of the conservation laws Eq. (9.77)

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

we have

$$\int T_{0\mu}\, \mathrm{d}^3y = \text{const.}, \tag{9.104}$$

i.e. the total energy and momentum are conserved. It follows that the $1/r$ part of $\bar{h}_{0\mu}$ is time independent to lowest order, so it will not contribute to any wave field. This generalizes Eq. (9.80). Again, see Exer. 32, § 9.7. Then, using Eq. (9.81), we get the generalization of Eq. (9.83):

$$\bar{h}_{jk}(t, x^i) = -\frac{2}{r} I_{jk,00}(t-r). \tag{9.105}$$

As before, we can adopt the TT gauge to get

$$\begin{aligned} \bar{h}^{\mathrm{TT}}_{xx} &= \frac{1}{r}[\not{I}_{xx,00}(t-r) - \not{I}_{yy,00}(t-r)], \\ \bar{h}^{\mathrm{TT}}_{xx} &= \frac{2}{r}[\not{I}_{xx,00}(t-r). \end{aligned} \tag{9.106}$$

9.4 The energy carried away by gravitational waves

Preview

We have seen that gravitational waves can put energy into things they pass through. This is how bar detectors work. It stands to reason, then, that they also carry energy away from their sources. This is a very important aspect of gravitational wave theory because, as we shall see, there are some circumstances in which the effects of this loss of energy on a source can be observed, even when the gravitational waves themselves cannot be detected. There are a number of different methods of deriving the formula for this energy loss (see Misner *et al.* 1973) and the problem has attracted a considerable amount of effort and has been attacked from many different points of view; see Damour (1987), Schutz (1980a), Futamase (1983), or Blanchet (2006). Our approach here will remain within the simple case of linear theory and will make the maximum use of what we already know about the waves.

In our discussion of the harmonic oscillator as a detector of waves in § 9.2, we implicitly assumed that the detector was a kind of 'test body', where the influence on the gravitational wave field is negligible. But this is, strictly speaking, inconsistent. If the detector acquires energy from the waves, then surely the waves must be weaker after passing through the detector. That is, 'downstream' of the detector they should have slightly lower amplitude than 'upstream'. It is easy to see how this comes about once we realize that in § 9.2 we ignored the fact that the oscillator, once set in motion by the waves, will radiate waves itself. We solved this in § 9.3 and found, in Eq. (9.88), that waves of two frequencies will be emitted. Consider the emitted waves with frequency Ω, the same as the incident wave. The part that is emitted exactly downstream has the same frequency as the incident wave, so the *total* downstream wave field has an amplitude that is the sum of the two. We will see below that the two interfere destructively, producing a net decrease in the downstream amplitude (see Fig. 9.5). (In other directions, there is no net interference: the waves simply pass through each other.) By assuming that this amplitude change signals a change in the energy actually carried by the waves, and by equating this energy change to the energy extracted from the waves by the detector, we will arrive at a simple expression for the energy carried by the waves as a function of their amplitude. We will then be able to calculate the energy lost by bodies that radiate arbitrarily, since we know from § 9.3 what waves they produce.

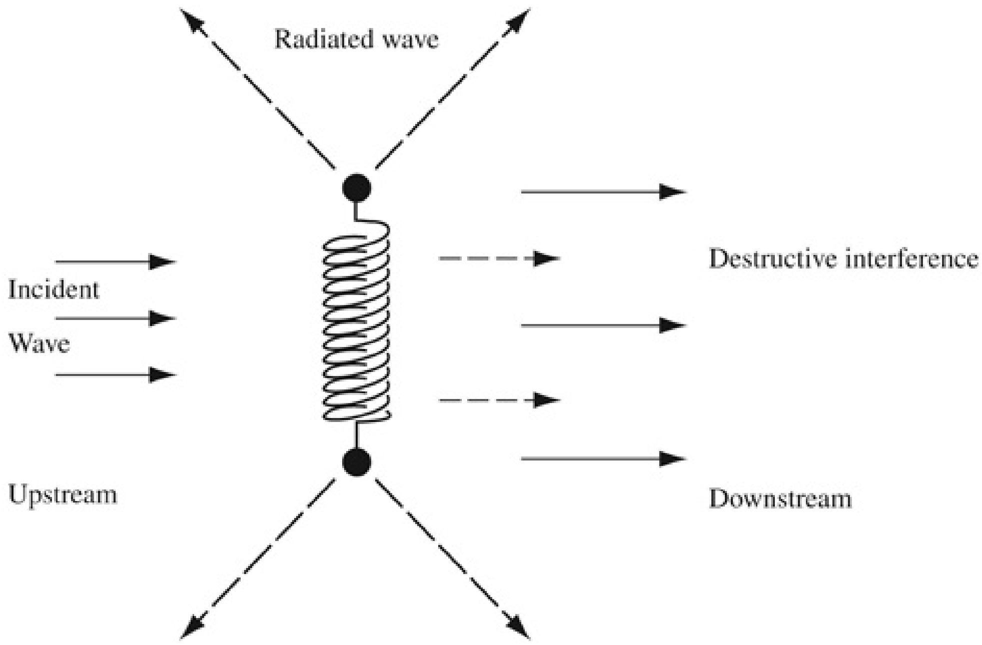


Figure 9.5 When the detector of Fig. 9.2 is excited by a wave, it re-radiates some waves itself.

The energy flux of a gravitational wave

What we are after is the energy flux, the energy carried by a wave across a surface per unit area per unit time. It is more convenient, therefore, to consider not just one oscillator but an array of them filling the plane $z = 0$. We suppose they are very close together, so we may regard them as a nearly continuous distribution of oscillators, σ oscillators per unit area (Fig. 9.6). If the incident wave is, in the TT gauge,

$$\begin{aligned} \bar{h}^{\mathrm{TT}}_{xx} &= A\cos\Omega(z-t), \\ \bar{h}^{\mathrm{TT}}_{yy} &= -\bar{h}^{TT}_{xx}, \end{aligned} \tag{9.107}$$

all other components vanishing, then in § 9.2 we have seen that each oscillator responds with a steady oscillation (after transients have died out) of the form

$$\xi = R\cos(\Omega t + \phi), \tag{9.108}$$

where R and ϕ are given by Eqs. (9.48) and (9.49) respectively. This motion is steady because the energy dissipated by friction in the oscillators is compensated by the work done on the spring by the tidal gravitational forces of the wave. It follows that the wave supplies an energy to each oscillator at a rate equal to

$$\mathrm{d}E/\mathrm{d}t = \nu\,(\mathrm{d}\xi/\mathrm{d}t)^2 = m\gamma\,(\mathrm{d}\xi/\mathrm{d}t)^2. \tag{9.109}$$

Averaging this over one period of oscillation, $2\pi/\Omega$, in order to get a steady energy loss, gives (angle brackets denote the average)

$$\begin{aligned}\langle \mathrm{d}E/\mathrm{d}t\rangle &= \frac{1}{2\pi/\Omega}\int_0^{2\pi/\Omega} m\gamma\Omega^2R^2\sin^2(\Omega t+\phi)\,\mathrm{d}t\\ &= \tfrac{1}{2}m\gamma\Omega^2R^2.\end{aligned} \tag{9.110}$$

This is the energy supplied to each oscillator per unit time. With σ oscillators per unit area, the net energy flux F of the wave must decrease on passing through the $z = 0$ plane by

$$\delta F = -\tfrac{1}{2}\sigma m\gamma\Omega^2R^2. \tag{9.111}$$

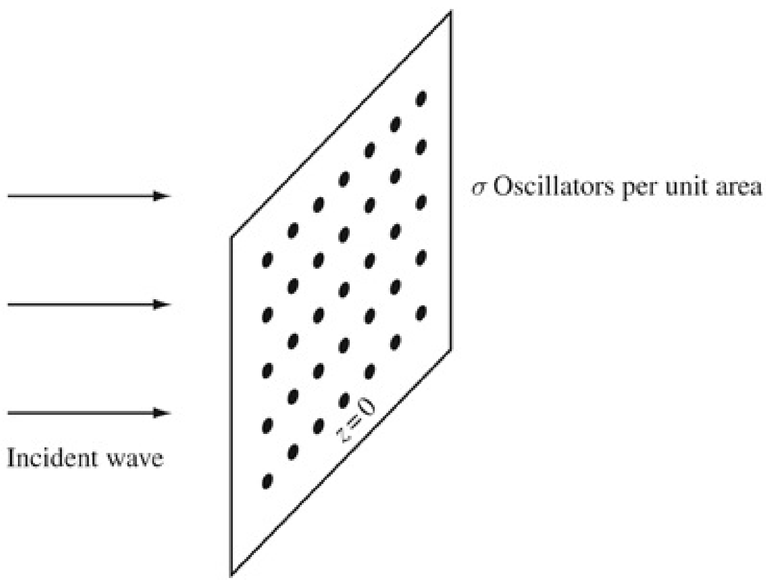


Figure 9.6 The situation when detectors of Fig. 9.5 are arranged in a plane at a density of σ per unit area.

We calculate the change in the amplitude downstream independently of the calculation that led to Eq. (9.111). Each oscillator has a quadrupole tensor given by Eq. (9.88), with ωt replaced by $\Omega t + \phi$ and A replaced by $R/2$. (Each mass moves an amplitude A, one-half of the total stretching of the spring R.) Since in our case R is tiny compared to $l_0(R = 0(h_{xx}^{\mathrm{TT}}\, l_0))$, the 2Ω term in Eq. (9.88) is negligible compared to the Ω term. So each oscillator has

(9.112)

$$I_{xx} = ml_0 R\cos(\Omega t + \phi). \tag{9.112}$$

By Eq. (9.83), each oscillator produces a wave amplitude

$$\delta\bar{h}_{xx} = -2\Omega^2 ml_0 R\cos[\Omega(r - t) - \phi]/r \tag{9.113}$$

at any point a distance r away. (We call it $\delta\bar{h}_{xx}$ to indicate that it is small compared to the incident wave.) It is a simple matter to get the total radiated field by adding up the contributions due to all the oscillators. In Fig. 9.7, consider a point P a distance z downstream from the plane of oscillators. Set up polar coordinates ($\tilde{\omega}$, ϕ) in the plane, centered at Q beneath P. A typical oscillator O at a distance $\tilde{\omega}$ from Q contributes a field, Eq. (9.113), at P, with $r = (\tilde{\omega}^2 + z^2)^{1/2}$. Since the number of such oscillators between $\tilde{\omega}$ and $\tilde{\omega} + \mathrm{d}\tilde{\omega}$ is $2\pi\sigma\tilde{\omega}\,\mathrm{d}\tilde{\omega}$, the total oscillator-produced field at P is

$$\delta\bar{h}_{xx}^{\text{total}} = -2m\Omega^2 l_0 R 2\pi \int_0^\infty \sigma\cos[\Omega(r - t) - \phi]\frac{\tilde{w}\,\mathrm{d}\tilde{w}}{r}.$$

But we may change the integration variable to r,

$$\tilde{w}\,\mathrm{d}\tilde{w} = r\,\mathrm{d}r,$$

obtaining

$$\delta \bar{h}_{xx}^{\text{total}} = -2m\Omega^2 l_0 R 2\pi \int_z^\infty \sigma \cos[\Omega(r-t)-\phi]\, \mathrm{d}r. \tag{9.114}$$

If σ were constant, this would be trivial to integrate, but its value would be undefined at $r = \infty$. Physically, we should expect that the distant oscillators play no real role, so we adopt the device of assuming that σ is proportional to exp $(-\varepsilon r)$ and allowing ε to tend to zero after the integration. The result is

$$\delta \bar{h}_{xx}^{\text{total}} = 4\pi \sigma m \Omega l_0 R \sin[\Omega(z-t)-\phi]. \tag{9.115}$$

So the plane of oscillators sends out a net plane wave. To compare this to the incident wave, we must put Eq. (9.115) in the same TT gauge (recall that Eq. (9.83) is *not* in the TT gauge), with the result (Exer. 42, § 9.7)

$$\delta \bar{h}_{xx}^{\text{TT}} = -\delta \bar{h}_{xx}^{\text{TT}} = 2\pi \sigma m \Omega l_0 R \sin[\Omega(z-t)-\phi]. \tag{9.116}$$

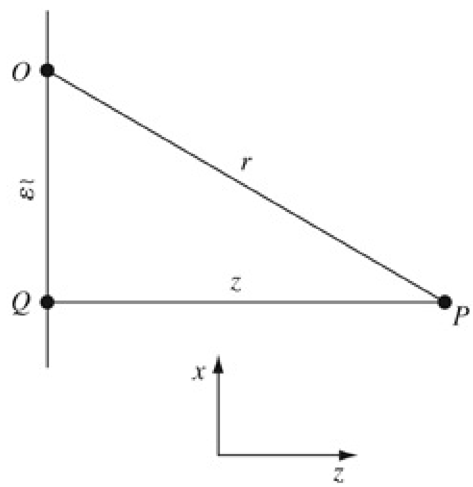


Figure 9.7 Geometry for calculating the field at *P* due to an oscillator at *O*.

If we now add this to the incident wave, Eq. (9.107), we get the net result, to first order in *R*,

$$\begin{aligned} \bar{h}_{xx}^{\text{net}} &= \bar{h}_{xx}^{\text{TT}} + \delta \bar{h}_{xx}^{\text{TT}} \\ &= (A - 2\pi \sigma m \Omega l_0 R \sin\phi) \cos[\Omega(z-t)-\psi], \end{aligned} \tag{9.117}$$

where

$$\tan\psi = \frac{2\pi\sigma m\Omega l_0 R}{A}\cos\phi. \tag{9.118}$$

Apart from a small phase shift ψ, the net effect is a reduction in the amplitude A by

$$\delta A = -2\pi\sigma m\Omega l_0 R\,\sin\phi. \tag{9.119}$$

This reduction must be responsible for the decrease in flux F downstream. Dividing Eq. (9.111) by Eq. (9.119) and using Eqs. (9.48) and (9.49) to eliminate R and ϕ gives the remarkably simple result

$$\frac{\delta F}{\delta A} = \frac{1}{16\pi}\Omega^2 A. \tag{9.120}$$

This is our key result. It says that a change δA in the amplitude A of a wave of frequency Ω changes its flux F (averaged over one period) by an amount depending only on Ω, A, and δA. The oscillators helped us to derive this result from conservation of energy, but they have dropped out completely! Eq. (9.120) is a property of the wave itself. We can 'integrate' Eq. (9.120) to get the total flux of a wave of frequency Ω and amplitude A:

$$F = \frac{1}{32\pi}\Omega^2 A^2. \tag{9.121}$$

Since the average of the square of the wave, Eq. (9.107), is

$$\langle(\bar{h}^{\rm TT}_{xx})^2\rangle = \tfrac{1}{2}A^2$$

(again, angle brackets denote an average over one period), and since there are only two nonvanishing components of $\bar{h}^{\rm TT}_{\mu\nu}$, we can also write Eq. (9.121) as

---

$$F = \tfrac{1}{32\pi}\Omega^2\langle\bar{h}^{\rm TT}_{\mu\nu}\bar{h}^{{\rm TT}\mu\nu}\rangle. \tag{9.122}$$

This form is invariant under background Lorentz transformations, but not under gauge changes. Since one polarization can be transformed into another by a background Lorentz transformation (a rotation), Eq. (9.122) applies to all polarizations and hence to arbitrary plane waves of frequency Ω. In fact, since it gives an energy rate per unit area, it applies to *any* wavefront, either plane waves or the spherical expanding ones of § 9.3: we can always look at a small enough area that the curvature of the wavefront is not noticeable. The generalization to arbitrary waves (no single frequency) is in Exer. 43, § 9.7.

The reader who remembers the discussion of energy in § 7.3 may object that this whole derivation is suspect because of the difficulty of defining energy in GR. Indeed, we have not *proved* that energy is conserved, that the energy put into the oscillators must equal the decrease in flux; we have simply assumed this in order to derive the flux. Our proof may be turned around, however, to argue that the flux we have constructed is the only acceptable definition of energy for the waves, since our calculation shows it is conserved, when added to other energies, to lowest order in $h_{\mu\nu}$. The qualification 'to lowest order' is important, since it is precisely because we are almost in flat spacetime that, at lowest order, we can construct conserved quantities. At higher order, away from linearized theory, local energy cannot be so easily defined, because the time dependence of the true metric becomes important. These questions are among the most fundamental in relativity, and are discussed in detail in any of the advanced texts. Our equations should be used *only* in linearized theory.

Energy lost by a radiating system

Consider a general isolated system, radiating according to Eqs. (9.82)–(9.87). By integrating Eq. (9.122) over a sphere surrounding the system, we can calculate its net energy loss rate. For example, at a distance r along the z axis, Eq. (9.122) is

$$F = \frac{\Omega^6}{32\pi r^2}\langle 2(\bar{I}_{xx} - \bar{I}_{yy})^2 + 8\bar{I}^{xy}\rangle. \tag{9.123}$$

Use of the identity

$$\bar{I}^i{}_i = \bar{I}_{xx} + \bar{I}_{yy} + \bar{I}_{zz} = 0 \tag{9.124}$$

(which follows from Eq. (9.87)) gives, after some algebra,

$$F = \frac{\Omega^6}{16\pi r^2} \langle 2\bar{I}_{ij}\bar{I}^{ij} - 4\bar{I}_{zj}\bar{I}_z{}^j + \bar{I}^2_{zz} \rangle. \tag{9.125}$$

Now, the index z appears here only because it is the direction from the center of the coordinates, where the radiation comes from. It is the only part of F which depends on the location on the sphere of radius r about the source, since all the components of $\bar{I}_{ij}$ depend on time but not position. Therefore we can generalize Eq. (9.125) to arbitrary locations on the sphere by using the unit vector normal to the sphere,

$$n^j = x^j/r. \tag{9.126}$$

We get for F

$$F = \frac{\Omega^6}{16\pi r^2} \langle 2\bar{I}_{ij}\bar{I}^{ij} - 4n^j n^k \bar{I}_{ji}\bar{I}_k{}^i + n^i n^j n^k n^l \bar{I}_{ij}\bar{I}_{kl} \rangle. \tag{9.127}$$

The total luminosity of the source is the integral of this over the sphere of radius r. In Exer. 45, § 9.7 we prove the following integrals over the entire sphere

$$\int n^j n^k \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi = \frac{4\pi}{3}\delta^{jk}, \tag{9.128}$$

$$\int n^i n^j n^k n^l \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi = \frac{4\pi}{15}(\delta^{ij}\delta^{kl} + \delta^{ik}\delta^{jl} + \delta^{il}\delta^{jk}). \tag{9.129}$$

It then follows that the luminosity L of a source of gravitational waves is

$$\begin{aligned} L = \int F r^2 \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi = \tfrac{1}{4}\Omega^6 \langle 2\bar{I}_{ij}\bar{I}^{ij} - \tfrac{4}{3}\bar{I}_{ij}\bar{I}^{ij} \\ + \tfrac{1}{15}(\bar{I}^i{}_i\bar{I}^k{}_k + \bar{I}^{ij}\bar{I}_{ij} + \bar{I}^{ij}\bar{I}_{ij}) \rangle, \end{aligned} \tag{9.130}$$

$$L = \frac{1}{5}\Omega^6 \langle \not{I}_{ij} \not{I}^{ij} \rangle. \tag{9.131}$$

The generalization to cases where $\not{I}_{ij}$ has a more general time dependence is

$$L = \frac{1}{5}\langle \dddot{\not{I}}_{ij} \dddot{\not{I}}^{ij} \rangle, \tag{9.132}$$

where dots denote time derivatives.

It must be stressed that Eqs. (9.121) and (9.132) are acccurate only for weak gravitational fields and slow velocities. They can at best give only order-of-magnitude results for highly relativistic sources of gravitational waves. But in the spirit of our derivation and discussion of the order-of-magnitude estimate of h_{ij} in Eq. (9.70), we can still learn something about strong sources from Eq. (9.121). Since I_{jk} is of order MR^2, Eq. (9.121) tells us that $L \sim M^2R^4\Omega^6 \sim (M/R)^2(R\Omega)^6 \sim \phi_0^2 v^6$. The luminosity is a very sensitive function of the velocity. The largest velocities we should expect are of the order of the velocity of free fall, $v^2 \sim \phi_0$, so we should expect

$$L \lesssim (\phi_0)^5. \tag{9.133}$$

Since $\phi_0 \lesssim 1$, the luminosity in geometrized units should never substantially exceed one. In ordinary units this is

$$L \lesssim 1 = c^5/G \approx 3.6 \times 10^{52}\,\mathrm{W}. \tag{9.134}$$

We can understand why this particular luminosity is an upper limit by the following simple argument. The radiation field inside a source of size R and luminosity L has energy density $\gtrsim L/R^2$ (because $|T^{0i}| \sim |v^i|T^{00} = cT^{00} = T^{00}$), which is the flux across its surface. The total energy in radiation is therefore $\gtrsim LR$. The Newtonian potential of the radiation alone is therefore $\gtrsim L$. We shall see in the next chapter that anything where the Newtonian potential substantially exceeds one must form a black hole: its gravitational field will be so strong that no radiation will escape at all. Therefore $L \sim 1$ is the largest luminosity any source can have. This argument applies equally well to all forms of radiation, electromagnetic as well as gravitational. The brightest quasars and gamma-ray bursts, which are the most luminous classes of object so far observed, have a (geometrized) luminosity $\lesssim 10^{-10}$. By contrast, black hole mergers (Ch. 11) have been shown by numerical simulations to reach peak luminosities $\sim 10^{-3}$, all of the energy of course emitted in gravitational waves.

An example. The Hulse –Taylor binary pulsar

In § 9.3 we calculated $I\!\!\!-_{ij}$ for a binary system consisting of two stars of equal mass M in circular orbits a distance l_0 apart. If we use the real part of Eq. (9.97) in Eq. (9.132), we get

$$L = \tfrac{8}{5} M^2 l_0^4 \omega^6. \tag{9.135}$$

Eliminating l_0 in favor of M and ω, we get

$$L = \frac{32}{5\sqrt[3]{4}} (M\omega)^{10/3} \approx 4.0 (M\omega)^{10/3}. \tag{9.136}$$

This expression illustrates two things: first, that L is dimensionless in geometrized units and, second, that it is almost always easier to compute in geometrized units, and then convert back at the end. The conversion is

$$\begin{aligned} L\,(\text{SI units}) &= \frac{c^5}{G} L\,(\text{geometrized}) \\ &= 3.63 \times 10^{52}\,\text{J s}^{-1} \times L\,(\text{geometrized}). \end{aligned} \tag{9.137}$$

So for the binary pulsar system described in § 9.3, if its orbit were circular, we would have $\omega = 2\pi/P = 7.5049 \times 10^{-13}\,\text{m}^{-1}$ and

$$L = 1.71 \times 10^{-29} \tag{9.138}$$

in geometrized units. We can, of course, convert this to watts, but a more meaningful procedure is to compare this with the Newtonian energy of the system, which is (defining the orbital radius $r = \tfrac{1}{2} l_0$),

$$
\begin{aligned}
E &= \tfrac{1}{2}M\omega^2 r^2 + \tfrac{1}{2}M\omega^2 r^2 - \frac{M^2}{2r} \\
&= \frac{M}{r}(\omega^2 r^3 - \tfrac{1}{2}M) = -\frac{M^2}{4r} \\
&= -4^{-2/3} M^{5/3}\omega^{2/3} \approx -0.40 M^{5/3}\omega^{2/3} && (9.139)\\
&= -1.11 \times 10^{-3}\,\mathrm{m}. && (9.140)
\end{aligned}
$$

The physical question is: How long does it take to change this? Put differently, the energy radiated in waves must change the orbit by decreasing its energy, which makes $|E|$ larger and hence ω larger and the period smaller. What change in the period do we expect in, say, one year?

From Eq. (9.139), by taking logarithms and differentiating, we get

$$
\frac{1}{E}\frac{\mathrm{d}E}{\mathrm{d}t} = \frac{2}{3}\frac{1}{\omega}\frac{\mathrm{d}\omega}{\mathrm{d}t} = -\frac{2}{3}\frac{1}{P}\frac{\mathrm{d}P}{\mathrm{d}t}. \tag{9.141}
$$

Since $\mathrm{d}E/\mathrm{d}t$ is just $-L$, we can solve for $\mathrm{d}P/\mathrm{d}t$:

$$
\begin{aligned}
\mathrm{d}P/\mathrm{d}t &= (3\,PL)/(2E) \approx -15\,\mathrm{P}M^{-1}(M\Omega)^{8/3} \\
&= -2.0 \times 10^{-13},
\end{aligned} \tag{9.142}
$$

which is dimensionless in any system of units. It can be reexpressed in seconds per year:

$$
\mathrm{d}P/\mathrm{d}t = -6.0 \times 10^{-6}\,\mathrm{s\,yr}^{1}. \tag{9.143}
$$

This estimate needs to be revised to allow for the eccentricity of the orbit, which is considerable: $e = 0.617$. The correct formula is derived in Exer. 49, § 9.7. The result is that the true rate of energy loss is some 12 times our estimate, Eq. (9.138). This is such a large factor because the stars' maximum angular velocity (when they are closest) is larger than the mean value we have used for Ω, and since L depends on the angular velocity to a high power, a small change in the angular velocity accounts for this rather large factor of 12. So the relativistic prediction is:

$$
\begin{aligned}
\mathrm{d}P/\mathrm{d}t &= -2.4 \times 10^{-12}, && (9.144)\\
&= -7.2 \times 10^{-5}\,\mathrm{s\,yr}^{-1}. && (9.145)
\end{aligned}
$$

The observed value as of 2004 is (Weisberg and Taylor 2005)

$$dP/dt = -(2.4184 \pm 0.0009) \times 10^{-12}. \tag{9.146}$$

The effect has been observed in other binaries as well (Lorimer 2008).

9.5 Astrophysical sources of gravitational waves

Overview

Physicists and astronomers have made great efforts to understand what kinds of sources of gravitational waves the current detectors might be able to see. This has been motivated partly by the need to decide whether the large investment in these detectors is justified, and partly because the sensitivity of the detectors depends on the accuracy with which waveforms can be predicted, so that they can be recognized against detector noise. While the astrophysics of potential sources is beyond the scope of this book, it is useful to review the basic classes of sources, learn what general relativity says about them, and understand why they might be interesting to observe. We shall consider four groups of sources: binary systems, spinning neutron stars, gravitational collapse, and the Big Bang.

Binary systems

We have seen how to compute the expected wave amplitude from a binary system in Eqs. (9.98) and (9.99), as well as in Exers. 29 and 39 in § 9.7. There are a number of known binary systems in our Galaxy which ought, by these equations, to be radiating gravitational waves in the frequency band observable by LISA, and with amplitudes well above LISA's expected instrumental noise. When LISA begins its observations, therefore, scientists will be looking for these signals as proof that general relativity is correct at this basic level, as well as that the spacecraft is operating properly.

The ground-based detectors will not, however, be looking for signals from long-lived binary systems. The reason is evident if we combine some of our previous computations. We have calculated what the gravitational wave luminosity of an equal-mass binary is (Eq. (9.136)) and what its binding energy is (Eq. (9.139)). We used the last two equations to derive the lifetime of the Hulse–Taylor binary pulsar. We can generalize this to find the lifetime of any equal-mass circular binary system, expressing it in terms of the masses M of the stars and the frequency of the orbit $f = \omega/2\pi$:

$$\tau_{\rm gw} = -\frac{E}{L} = 7.44 \times 10^{-4} M^{-5/3} f^{-8/3}, \tag{9.147}$$

$$= 2.43 \left(\frac{M}{\mathrm{M}_\odot}\right)^{-5/3} \left(\frac{f}{100\,\mathrm{Hz}}\right)^{-8/3} \ \mathrm{s}. \tag{9.148}$$

What the second equation tells us is that, if LIGO observes a binary system composed of stars of solar mass, it has only a few seconds to make the observation. During this time, the signal increases in frequency, something the gravitational wave scientists call a 'chirp'. There are no realistic systems in the frequency range of ground-based detectors that are long-lived. Instead, these detectors look for *inspiral* events ending in the merger of the two objects, which itself might produce a burst of radiation. For neutron stars, the merger happens when the signal frequency reaches about 2 kHz, a number that is sensitive to the neutron star equation of state (see the next chapter). If the objects in the binary are black holes of mass 10 $\mathrm{M}_\odot$, the final frequency is similar, and it scales inversely with the masses. Advanced LIGO and VIRGO should see several neutron-star mergers per year, and while the event rate for black holes is harder to predict, it is likely to be similar.

It is particularly interesting from the point of view of general relativity to observe the merger of two black holes. This can be simulated numerically, and by comparing the observed and predicted waveforms we have a unique test of general relativity in the strongest possible gravitational fields. This is also the only direct way to observe a black hole: after a merger of black holes or neutron stars has led to a single black hole, that hole will oscillate for a short time until is radiates away all its deformities and settles down as a smooth Kerr black hole (see Ch. 11). This 'ringdown radiation' carries a distinctive signature that will distinguish the black hole from any neutron star or other material system.

LISA, observing between 0.1 mHz and 10 mHz, will follow the coalescence and merger of black holes around $10^6 M_\odot$. Astronomers know that such black holes exist in the centers of most galaxies, including our own Milky Way (Merritt and Milosavljevic 2005), as we discuss in Ch. 11. LISA will have sufficient sensitivity to see such mergers anywhere in the universe, even back to the time of the formation of the first stars and galaxies. Its observations may be very informative about the way galaxies themselves formed and merged in the early universe.

Notice that, if we observe a chirp signal well enough to measure its inspiral timescale τ from the rate of change of the signal's frequency, then we can infer the mass M of the system from Eq. (9.147). LISA by itself, or a network of ground-based detectors working together, can measure the degree of circular polarization of the waves. This contains the information about the inclination of the orbit to the line of sight, and allows us to compute from the observed amplitude of the waves some standard amplitude, such as the amplitude the same system would be radiating if it were oriented face-on. Then we can go back to Eq. (9.98), eliminate l_0 in favor of the mass and frequency, and be left with just one unknown: the distance r to the source. Remarkably, this property holds even if the binary system does not have equal masses: chirping binary signals contain enough information to deduce the distance to the source (Schutz 1986). Gravitational wave astronomers call these systems 'standard sirens', by analogy with the usual standard candles of optical astronomy, which we will discuss in Ch. 12. We will see there that gravitational wave observations of black-hole binaries by LISA may assist astronomers measure the large-scale dynamics of the universe.

Spinning neutron stars

Neutron stars are very compact stars formed in gravitational collapse. We will study them as relativistic stellar objects in the next chapter. Here we simply note that many neutron stars are pulsars, whose spin sweeps a beam of electromagnetic radiation past the Earth each time they turn. Many spin rapidly, at frequencies above 20 Hz, and if these radiate gravitational waves, then they would be in the observing band of ground-based detectors. There could in principle be many stars not known as pulsars that also spin this rapidly, because their beams do not cross the Earth. Moreover, radio surveys for pulsars only cover the near neighborhood of the Sun in our Galaxy; there could be more distant pulsars that are not yet known.

Such stars could radiate gravitational waves if they are not symmetric about the rotation axis. Pulsars are clearly not symmetric, since they beam their radiation somehow. But it is not clear how much mass asymmetry is required to produce the beaming. Other asymmetries could come from frozen-in irregularities in the semi-solid outer layer of a neutron star (called its ‘crust’), or in a possible solid core. It is also known that spinning neutron stars are vulnerable to a gravitational-wave driven instability called the r-mode instability, which could produce significant radiation. We can compute the radiation due to mass asymmetry from Eqs. (9.84)–(9.86). If the star is nearly axisymmetric, then we can approximate the amplitude of either of the polarizations radiated along the spin axis by the formula

$$h \sim 2\varepsilon\Omega^2 I_{\mathrm{NS}}/r,$$

where I_{NS} is the moment of inertia of the spherical neutron star and ε is the fractional asymmetry of the star about the spin axis. If we use typical values of $I_{\mathrm{NS}} = 10^{38}$ kg m^2, $r = 1$ kpc (about 3×10^{22} m), $\Omega = 2\pi f$ with $f = 60$ Hz, and $\varepsilon = 10^{-5}$, then we get $h \sim 10^{-25}$. This is a very small amplitude, but not impossibly small. Scientists find such small signals by taking long stretches of data and filtering for them, essentially by performing a Fourier transform. The Fourier transform concentrates the power of the signal in one frequency band, while distributing the noise power of the data stream over the whole observing band. To go from the Advanced LIGO broad-band sensitivity of around 10^{-22} to a sensitivity of 10^{-25} for a narrow-band signal like the one we are considering here, the data analyst must have taken a number of cycles of the waveform equal to at least the square of the ratio of these two numbers, or 10^6. For this frequency, this would take less than a day.

Until a spinning neutron star has been observed, we won't know what a reasonable value for ε is. However, for many known pulsars we can already set limits from radio observations. This is because, as for the binaries we considered above, a radiating pulsar loses energy. This causes it to spin down. Most pulsars are observed to spin down, and so their observed slowing rate sets an upper limit on the possible amplitude of gravitational waves. It is only an upper limit, because it is very likely that the spindown is dominated by other effects, such as losses to electromagnetic radiation and particle emission, so that gravitational waves play a minor role. For known pulsars the limits obtained on ε this way range from 10^{-3} to below 10^{-7}.

It is therefore desirable to do searches for gravitational wave pulsars using months rather than days of data. When the pulsar's position and frequency are known from radio observations, this is not a difficulty, but when gravitational wave astronomers try to search the entire sky for unknown neutron stars, the computational demands become enormous. This is because the apparent frequency of the pulsar signal is strongly Doppler modulated by the Earth's spin and orbital motion during a period as long as a month or more, and the details of the modulation depend on the star's location on the sky. Data analysts therefore have to search many different locations separately to perform their filtering. At present, this is a problem that would overwhelm the most powerful computers in existence. The gravitational wave projects are getting help in this analysis from the general public, using the screen-saver called Einstein@Home.

Gravitational collapse

The objective that motivated Joseph Weber to develop the first bar detector was to register waves from a supernova. The spectacular optical display of a supernova explosion masks what really happens inside: the compact core of a giant star, having exhausted its supply of energy from nuclear reactions, collapses inward, and the subsequent dynamics can convert some of the energy released into the explosion that blows off the envelope of the star. But what happens to the collapsing core, and how that energy is converted into the explosion, is not well understood because it is impossible to observe the core directly. Gravitational waves, along with neutrinos, provide the only probes that come to us directly from the core.

The amplitude of gravitational waves to be expected is very uncertain. It is sensitive to the initial state of rotation of the core, to instabilities that develop during collapse, and to poorly understood details of the physics of dense matter. Modeling collapse on a computer is difficult, and the predictions so far are only approximate. However, there is wide agreement that the amplitudes are likely to be far smaller than the order-of-magnitude estimate we made in the opening paragraph of § 9.3.

What is more, it is not possible at present to predict a detailed waveform, so that the data analysts cannot dig so deeply into the noise of the detector as they can for binaries or spinning neutron stars. All of these circumstances make it seem less likely that the first detected signal will be that of a supernova explosion. That expectation would be reversed, however, if the Galaxy experienced another supernova explosion like SN1987a, which occurred in its satellite galaxy, the Large Magellanic Cloud.

Gravitational waves from the Big Bang

The study of the large-scale structure of the universe, and its history, is called cosmology, and it will be the subject of Ch. 12 below. Cosmology has undergone a revolution since the 1980s, with a huge increase in data and in our insight into what went on in the early universe. Part of that revolution impacts on the study of gravitational waves: it seems very probable that the very early universe was the source of a random sea of gravitational radiation that even today forms a background to our observations of other sources.

The radiation originated in a host of individual events too numerous to count.

The waves, superimposed now, have very similar character to the random noise that comes from instrumental effects. Although the radiation was intense when it was generated, the expansion of the universe has cooled it down, and one of the most uncertain aspects of our understanding is what intensity it should have today. It is possible that it will be strong enough that, as detectors improve their sensitivity, they will encounter a 'noise' that does not go away, and that can be shown to be isotropic on the sky. In exactly this way, Penzias (1979) and Wilson (1979) discovered the cosmic microwave background radiation in a radio receiver at Bell Labs, an event for which they were awarded the Nobel Prize for Physics.

However, it is more likely that the radiation is weaker and will remain below the noise in our detectors for some time to come. How, then, can we find it? The answer is that, while it is a random noise in any one detector, the randomness is correlated between detectors. Two detectors in the same place experience exactly the same noise. If we make a correlation of their output (simply multiplying them and integrating in time) we should obtain a nonzero result much larger than we expect from the variance of the correlation of two statistically independent noise fields. In practice, the most sensitive pairs of detectors are the two LIGO installations, and the VIRGO-GEO600 pair. Both have separations between them so that the correlations in their random wavefields would not be perfect. However, gravitational waves with wavelengths longer than the separation will still be well correlated, and this allows these detectors to search for a background.

At present, the only limits we have are from the two LIGO detectors, and they are not surprising. Cosmologists express the strength of backgrounds in terms of the energy density they carry, as a fraction of the total energy density of all the material in the universe, averaged over large volumes. We know from present observations that the energy density in random waves in the LIGO observing band is not larger than a fraction 10^{-5} of the total. It is hoped that Advanced LIGO may approach a limit around 10^{-10} of the total.

LISA can also make observations of the background. In its case, the background would have to be stronger than instrumental noise: correlation gains it nothing. But LISA's sensitivity in its waveband is great, and it seems likely that it would be able to detect a background around 10^{-10} of the total. Observations of the cosmic microwave background could also detect this radiation, at very low frequencies.

Pulsar timing might be able to detect a random background of gravitational waves from astrophysical systems, but it seems likely that these backgrounds will be larger than the cosmological background. This is a general problem, and it may be that only at frequencies above about 0.1 Hz will the universe be quiet enough to allow us to listen directly to the hiss of gravitational waves from the Big Bang.

Theoretical predictions of the radiation to be expected vary hugely, from below 10^{-15} up to 10^{-8} and higher. This reflects the uncertainty in theoretical models of the physical conditions and indeed of the laws of physics themselves during the early Big Bang, and demonstrates the importance that detecting a background would play in constraining these models. An observation of the random (stochastic) background of gravitational waves is possibly the most important observation that gravitational wave detectors can make.

9.6 Further reading

Joseph Weber's early thinking about detectors is in Weber (1961). One of the most interesting theoretical developments stimulated by research into gravitational wave detection has been the design of so-called 'quantum nondemolition' detectors: methods of measuring aspects of the excitation of a vibrating bar to arbitrary precision without disturbing the quantity being measured, even when the bar is excited only at the one-or two-quantum (phonon) level. See early work by Thorne *et al.* (1979) and Caves *et al.* (1980).

A full discussion of the wave equation is beyond our scope here, but is amply treated in many texts on electromagnetism, such as Jackson (1975). A simplified discussion of gravitational waves is in Schutz (1984). See also Schutz and Ricci (2001).

The detection of gravitational waves is a rapidly evolving field, so the student who wants the latest picture should consult the literature, starting with the various articles that we have cited from the open-access electronic journal *Living Reviews in Relativity*, whose review articles are kept up-to-date: Armstrong (2006), Blanchet (2006), Futamase and Itoh (2007), Hough and Rowan (2000), Will (2006). More popular-style articles about gravitational waves and other applications of general relativity can be found on the Einstein Online website: `http://www.einstein-online.info/en/`.

The websites of the detectors LIGO (`http://www.ligo.caltech.edu/`), GEO (`http://geo600.aei.mpg.de/`), LSC (`http://www.ligo.org/`), VIRGO (`http://wwwcascina.virgo.infn.it/`), and LISA (http://www.lisascience.org/) are also good sources of current information. The bar detectors of the Rome group (http://www.roma1.infn.it/rog/) and the Auriga detector (http://www.auriga.lnl.infn.it/) are the last operating resonant-mass detectors.

Readers who wish to assist with the compute-intensive analysis of data from the big interferometers may download a screen-saver called Einstein@Home, which uses the idle time on a computer to perform parts of the data analysis. Hundreds of thousands of computers have so far joined this activity. See the website http://einstein.phys.uwm.edu/.

9.7 Exercises

A function $f(s)$ has derivative $f'(s) = \mathrm{d}f/\mathrm{d}s$. Prove that $\partial f(k_\mu x^\mu)/\partial x^\nu = k_\nu f'(k_\mu x^\mu)$. Use this to prove Eq. (9.4) and the one following it.

Show that the real and imaginary parts of Eq. (9.2) at a fixed spatial position $\{x^i\}$ oscillate sinusoidally in time with frequency $\omega = k^0$.

Let $\bar{h}^{\alpha\beta}(t, x^i)$ be any solution of Eq. (9.1) that has the property $\int \mathrm{d}x^\alpha |\bar{h}^{\mu\nu}|^2 < \infty$, for the integral over any particular x^α holding other coordinates fixed. Define the Fourier transform of $\bar{h}^{\alpha\beta}(t, x^i)$ as

$$\bar{H}^{\alpha\beta}(\omega, k^i) = \int \bar{h}^{\alpha\beta}(t, x^i) \exp\left(\mathrm{i}\omega t - \mathrm{i}k_j x^j\right) \mathrm{d}t\, \mathrm{d}^3x.$$

Show, by transforming Eq. (9.1), that $\bar{H}^{\alpha\beta}(\omega, k^i)$ is zero except for those values of ω and k^i that satisfy Eq. (9.10). By applying the inverse transform, write $\bar{h}^{\alpha\beta}(t, x^i)$ as a superposition of plane waves.

Derive Eqs. (9.16) and (9.17).

(a) Show that $A^{(\mathrm{NEW})}_{\alpha\beta}$, given by Eq. (9.17), satisfies the gauge condition $A^{\alpha\beta}k_\beta = 0$ if $A^{(\mathrm{OLD})}_{\alpha\beta}$ does.

(b) Use Eq. (9.18) for $A^{(\mathrm{NEW})}_{\alpha\beta}$ to constrain B^μ.

(c) Show that Eq. (9.19) for $A^{(\mathrm{NEW})}_{\alpha\beta}$ imposes only three constraints on B^μ, not the four that we might expect from the fact that the free index α can take any values from 0 to 3. Do this by showing that the particular linear combination $k^\alpha(A_{\alpha\beta}U^\beta)$ vanishes for any B^μ.

(d) Using (b) and (c), solve for B^μ as a function of k^μ, $A^{(\mathrm{OLD})}_{\alpha\beta}$, and U^μ. These determine B^μ: there is no further gauge freedom.

(e) Show that it is possible to choose ξ^{β} in Eq. (9.15) to make any superposition of plane waves satisfy Eqs. (9.18) and (9.19), so that these are generally applicable to gravitational waves of any sort.

(f) Show that we cannot achieve Eqs. (9.18) and (9.19) for a static solution, *i.e.* one for which $\omega = 0$.

Fill in all the algebra implicit in the paragraph leading to Eq. (9.21).

Give a more rigorous proof that Eqs. (9.22) and (9.23) imply that a free particle initially at rest in the TT gauge remains at rest.

Does the free particle of the discussion following Eq. (9.23) *feel* any acceleration? For example, if the particle is a bowl of soup (whose diameter is much less than a wavelength), does the soup slosh about in the bowl as the wave passes?

Does the free particle of the discussion following Eq. (9.23) *see* any acceleration? To answer this, consider the two particles whose relative proper distance is calculated in Eq. (9.24). Let the one at the origin send a beam of light towards the other, and let it be reflected by the other and received back at the origin. Calculate the amount of proper time elapsed at the origin between the emission and reception of the light (you may assume that the particles' separation is much less than a wavelength of the gravitational wave). By monitoring changes in this time, the particle at the origin can 'see' the relative acceleration of the two particles.

(a) We have seen that

$$h_{yz} = A \sin \omega(t - x), \text{all other } h_{\mu\nu} = 0,$$

with A and ω constants, $|A| \ll 1$, is a solution to Eqs. (9.1) and (9.11). For this metric tensor, compute all the components of $R_{\alpha\beta\mu\nu}$ and show that some are not zero, so that the spacetime is not flat.

(b) Another metric is given by

$$h_{yz} = A \sin \omega(t - x), \quad h_{tt} = 2B(x - t),$$
$$h_{tx} = -B(x - t), \quad \text{all other } h_{\mu\nu} = 0.$$

Show that this also satisfies the field equations and the gauge conditions.

(c) For the metric in (b), compute $R_{\alpha\beta\mu\nu}$. Show that it is the *same* as in (a).

(d) From (c) we conclude that the geometries are identical, and that the difference in the metrics is due to a small coordinate change. Find a ξ^{μ} such that

$$h_{\mu\nu}(\text{part } a) - h_{\mu\nu}(\text{part } b) = -\xi_{\mu,\nu} - \xi_{\nu,\mu}.$$

(a) Derive Eq. (9.27).

(b) Solve Eqs. (9.28a) and (9.28b) for the motion of the test particles in the polarization rings shown in Fig. 9.1.

Do calculations analogous to those leading to Eqs. (9.28) and (9.32) to show that the separation of particles in the *z* direction (the direction of travel of the wave) is unaffected.

One kind of background Lorentz transformation is a simple 45° rotation of the *x* and *y* axes in the $x - y$ plane. Show that under such a rotation from (x, y) to (x', y'), we have $h^{TT}_{x'y'} = -h^{TT}_{xx}, h^{TT}_{x'x'} = h^{TT}_{xy}$. This is consistent with Fig. 9.1.

) Show that a plane wave with $A_{xy} = 0$ in Eq. (9.21) has the metric

$$ds^2 = -dt^2 + (1 + h_+)dx^2 + (1 - h_+)dy^2 + dz^2, \tag{9.149}$$

where $h_+ = A_{xx} \sin[\omega(t - z)]$.

(b) Show that this wave does not change proper separations of free particles if they are aligned along a line bisecting the angle between the *x*- and *y*-axes.

(c) Show that a plane wave with $A_{xx} = 0$ in Eq. (9.21) has the metric

$$ds^2 = -dt^2 + dx^2 + 2h_\times dx\, dy + dy^2 + dz^2, \tag{9.150}$$

where $h_\times = A_{xy} \sin[\omega(t - z)]$.

(d) Show that the wave in (c) does not change proper separations of free particles if they are aligned along the coordinate axes.

(e) Show that the wave in (c) produces an elliptical distortion of the circle that is rotated by 45° to that of the wave in (a).

ı) A wave is said to be circularly polarized in the $x - y$ plane if $h^{TT}_{yy} = -h^{TT}_{xx}$ and $h^{TT}_{xy} = \pm i h^{TT}_{xx}$. Show that for such a wave, the ellipse in Fig. 9.1 rotates without changing shape.

(b) A wave is said to be elliptically polarized with principal axes *x* and *y* if $h^{TT}_{xy} = \pm i a h^{TT}_{xx}$, where *a* is some real number, and $h^{TT}_{yy} = -h^{TT}_{xx}$. Show that if $h^{TT}_{xy} = \alpha\, h^{TT}_{xx}$, where α is a complex number (the general case for a plane wave), new axes x' and y' can be found for which the wave is elliptically polarized with principal axes x' and y'. Show that circular and linear polarization are special cases of elliptical.

Two plane waves with TT amplitudes, $A^{\mu\nu}$ and $B^{\mu\nu}$, are said to have orthogonal polarizations if $(A^{\mu\nu})^* B_{\mu\nu} = 0$, where $(A^{\mu\nu})^*$ is the complex conjugate of $A^{\mu\nu}$. Show that if $A^{\mu\nu}$ and $B^{\mu\nu}$ are orthogonal polarizations, a 45° rotation of $B^{\mu\nu}$ makes it proportional to $A^{\mu\nu}$.

Find the transformation from the coordinates (t, x, y, z) of Eqs. (9.33)–(9.36) to the local inertial frame of Eq. (9.37). Use this to verify Eq. (9.38).

Prove Eq. (9.39).

Use the sum of Eqs. (9.40) and (9.41) to show that the center of mass of the spring remains at rest as the wave passes.

Derive Eq. (9.44) from Eq. (9.43), and then prove Eq. (9.45).

Generalize Eq. (9.45) to the case of a plane wave with arbitrary elliptical polarization (Exer. 15) traveling in an arbitrary direction relative to the separation of the masses.

Consider the equation of geodesic deviation, Eq. (6.87), from the point of view of the geodesic at the center of mass of the detector of Eq. (9.45). Show that the vector ξ as we have defined it in Eq. (9.42) is twice the connecting vector from the center of mass to one of the masses, as defined in Eq. (6.83). Show that the tidal force as measured by the center of mass leads directly to Eq. (9.45).

Derive Eqs. (9.48) and (9.49), and derive the general solution of Eq. (9.45) for arbitrary initial data at $t = 0$, given Eq. (9.46).

Prove Eq. (9.53).

Derive Eq. (9.56) from the given definition of Q.

) Use the metric for a plane wave with ‘+’ polarization, Eq. (9.58), to show that the square of the coordinate speed (in the TT coordinate system) of a photon moving in the x-direction is

$$\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2 = \frac{1}{1 + h_+}.$$

This is not identically one. Does this violate relativity? Why or why not?

(b) Imagine that an experimenter at the center of the circle of particles in Fig. 9.1 sends a photon to the particle on the circle at coordinate location $x = L$ on the positive-x axis, and that the photon is reflected when it reaches the particle and returns to the experimenter. Suppose further that this takes such a short time that h_+ does not change significantly during the experiment. To first order in h_+, show that the experimenter’s proper time that elapses between sending out the photon and receiving it back is $(2 + h_+)L$.

(c) The experimenter says that this proves that the proper distance between herself and the particle is $(1 + h_+/2)L$. Is this a correct interpretation of her experiment? If the experimenter uses an alternative measuring process for proper distance, such as laying out a number of standard meter sticks between her location and the particle, would that produce the same answer? Why or why not?

(d) Show that if the experimenter simultaneously does the same experiment with a particle on the y-axis at $y = L$, that photon will return after a proper time of $(2 - h_+)L$.

(e) The difference in these return times is $2h_+L$ and can be used to measure the wave's amplitude. Does this result depend on our use of TT gauge, *i.e.* would we have obtained the same answer had we used a different coordinate system?

ı) Derive the full three-term return relation, Eq. (9.63), for the rate of change of the return time for a beam traveling through a plane wave h_+ along the x-direction, when the wave is moving at an angle θ to the z-axis in the $x - z$ plane.

(b) Show that, in the limit where L is small compared to a wavelength of the gravitational wave, the derivative of the return time is the derivative of $t + \delta L$, where $\delta L = L \cos^2 \theta \, h(t)$ is the excess proper distance for small L. Explain where the factor of $\cos^2 \theta$ comes from.

(c) Examine the limit of the three-term formula in (a) when the gravitational wave is traveling along the x-axis too ($\theta = \pm\pi/2$): what happens to light going parallel to a gravitational wave?

ı) Reconstruct $\bar{h}_{\mu\nu}$ as in Eq. (9.66), using Eq. (9.68), and show that surfaces of constant phase of the wave move outwards for the $A_{\mu\nu}$ term and inwards for $Z_{\mu\nu}$.

(b) Fill in the missing algebra in Eqs. (9.69)–(9.71).

Eq. (9.67) in the vacuum region outside the source – *i.e.* where $S_{\mu\nu} = 0$ – can be solved by separation of variables. Assume a solution for $\bar{h}_{\mu\nu}$ has the form $\sum_{lm} A^{lm}_{\mu\nu} f_l(r) Y_{lm}(\theta, \phi)/\sqrt{r}$, where Y_{lm} is the spherical harmonic.

(a) Show that $f_l(r)$ satisfies the equation

$$f_l'' + \frac{1}{r} f'_l + \left[\Omega^2 - \frac{(l + \frac{1}{2})^2}{r^2}\right] f_l = 0.$$

(b) Show that the most general spherically symmetric solution is given by Eq. (9.68).

(c) Substitute the variable $s = \Omega r$ to show that f_l satisfies the equation

$$s^2\frac{d^2f_l}{ds^2}+s\frac{df_l}{ds}+[s^2-(l+\tfrac{1}{2})^2]f_l=0. \tag{9.151}$$

This is known as Bessel's equation, whose solutions are called Bessel functions of order $l+\frac{1}{2}$. Their properties are explored in most text-books on mathematical physics.

(d) Show, by substitution into Eq. (9.151), that the function $f_l/\sqrt{s}$ is a linear combination of what are called the spherical Bessel and spherical Neumann functions

$$j_l(s)=(-1)^l s^l\left(\frac{1}{s}\frac{d}{ds}\right)^l\left(\frac{\sin s}{s}\right), \tag{9.152}$$

$$n_l(s)=(-1)^{l+1} s^l\left(\frac{1}{s}\frac{d}{ds}\right)^l\left(\frac{\cos s}{s}\right). \tag{9.153}$$

(e) Use Eqs. (9.152) and (9.153) to show that for $s \gg l$, the dominant behavior of j_l and n_l is

$$j_l(s)\sim\frac{1}{s}\sin\left(s-\frac{l\pi}{2}\right), \tag{9.154}$$

$$n_l(s)\sim-\frac{1}{s}\cos\left(s-\frac{l\pi}{2}\right). \tag{9.155}$$

(f) Similarly, show that for $s \ll l$, the dominant behavior is

$$j_l(s)\sim s^l/(2l+1)!!, \tag{9.156}$$

$$n_l(s)\sim-(2l-1)!!/s^{l+1}, \tag{9.157}$$

where we use the standard double factorial notation

$$(m)!!=m(m-2)(m-4)\cdots 3\cdot 1 \tag{9.158}$$

for odd m.

(g) Show from (e) that the outgoing-wave vacuum solution of Eq. (9.67), for any fixed l and m, is

$$(\bar{h}_{\mu\nu})_{lm}=A^{lm}_{\mu\nu}h^{(1)}_l(\Omega r)\,e^{-i\Omega t}Y_{lm}(\theta,\phi), \tag{9.159}$$

where $h_l^{(1)}(\Omega r)$ is called the spherical Hankel function of the first kind,

$$h_l^{(1)}(\Omega r) = j_l(\Omega r) + \mathrm{i} n_l(\Omega r). \tag{9.160}$$

(h) Repeat the calculation of Eqs. (9.69)–(9.74), only this time multiply Eq. (9.67) by $j_l(r\Omega)Y^*_{lm}(\theta,\phi)$ before performing the integrals. Show that the left-hand side of Eq. (9.67) becomes, when so integrated, exactly

$$\varepsilon^2 \left(j_l(\Omega\varepsilon)\frac{\mathrm{d}}{\mathrm{d}r}B_{\mu\nu}(\varepsilon) - B_{\mu\nu}(\varepsilon)\frac{\mathrm{d}}{\mathrm{d}r}j_l(\Omega\varepsilon) \right),$$

and that when $\Omega\varepsilon \ll l$ this becomes (with the help of Eqs. (9.159) and (9.156)–(9.157) above, since we assume $r = \varepsilon$ is outside the source) simply $\mathrm{i}A^{lm}_{\mu\nu}/\Omega$. Similarly, show that the right-hand side of Eq. (9.67) integrates to $-16\pi\Omega^l \int T_{\mu\nu} r^l Y^*_{lm}(\theta,\phi)\mathrm{d}^3x/(2l+1)!\,!$ in the same approximation.

(i) Show, then, that the solution is Eq. (9.159), with

$$A^{lm}_{\mu\nu} = 16\pi \mathrm{i}\Omega^{l+1} J^{lm}_{\mu\nu}/(2l+1)!\,!, \tag{9.161}$$

where

$$J_{\mu\nu}^{lm} = \int T_{\mu\nu} r^l Y_{lm}^*(\theta, \phi) \, d^3x. \tag{9.162}$$

(j) Let $l = 0$ and deduce Eq. (9.73) and (9.74).

(k) Show that if $J_{\mu\nu}^{lm} \neq 0$ for some l, then the terms neglected in Eq. (9.161), because of the approximation $\Omega\varepsilon \ll 1$, are of the same order as the dominant terms in Eq. (9.161) for $l + 1$. In particular, this means that if $J_{\mu\nu} \neq 0$ in Eq. (9.72), any attempt to get a more accurate answer than Eq. (9.74) must take into account not only the terms for $l > 0$ but also neglected terms in the derivation of Eq. (9.74), such as Eq. (9.69).

Re-write Eq. (9.82a) for a set of N discrete point particles, where the masses are $\{m_{(A)}, A = 1, \ldots, N\}$ and the positions are $\{x^i_{(A)}\}$.

Calculate the quadrupole tensor I_{jk} and its traceless counterpart $\text{I\kern-0.5em--}_{jk}$ (Eq. (9.87)) for the following mass distributions.

(a) A spherical star where density is $\rho(r, t)$. Take the origin of the coordinates in Eq. (9.82) to be the center of the star.

(b) The star in (a), but with the origin of the coordinates at an arbitrary point.

(c) An ellipsoid of uniform density ρ and semiaxes of length a, b, c oriented along the x, y, and z axes respectively. Take the origin to be at the center of the ellipsoid.

(d) The ellipsoid in (c), but rotating about the z axis with angular velocity ω.

(e) Four masses m located respectively at the points $(a, 0, 0)$, $(0, a, 0)$, $(-a, 0, 0)$, $(0, -a, 0)$.

(f) The masses as in (e), but all moving counter-clockwise about the z axis on a circle of radius a with angular velocity ω.

(g) Two masses m connected by a massless spring, each oscillating on the x axis with angular frequency ω and amplitude A about mean equilibrium positions a distance l_0 apart, keeping their center of mass fixed.

(h) Unequal masses m and M connected by a spring of spring constant k and equilibrium length l_0, oscillating (with their center of mass fixed) at the natural frequency of the system, with amplitude $2A$ (this is the total stretching of the spring). Their separation is along the x axis.

This exercise develops the TT gauge for spherical waves.

(a) In order to transform Eq. (9.83) to the TT gauge, use a gauge transformation generated by a vector $\xi^\alpha = B^\alpha(x^\mu)e^{i\Omega(r-t)}/r$, where B^α is a slowly varying function of x^μ. Find the general transformation law to order $1/r$.

(b) Demand that the new $\bar{h}_{\alpha\beta}$ satisfy three conditions to order $1/r$: $\bar{h}_{0\mu} = 0$, $\bar{h}^\alpha{}_\alpha = 0$, and $\bar{h}_{\mu j}n^j = 0$, where $n^j := x^j/r$ is the unit vector in the radial direction. Show that it is possible to find functions B^α, which accomplish such a transformation *and* which satisfy $\square\, \xi^\alpha = 0$ to order $1/r$.

(c) Show that Eqs. (9.84)–(9.87) hold in the TT gauge.

(d) By expanding R in Eq. (9.103) but discarding r^{-1} terms, show that the higher-order parts of $\bar{h}_{0\mu}$ that are not eliminated by Eq. (9.104) are gauge terms to order v^2, *i.e.* up to second time derivatives in the expansion of $\bar{h}_{00}$ and first time derivatives in $\bar{h}_{0j}$ in Eq. (9.103).

(a) Let n^j be a unit vector in three-dimensional Euclidean space. Show that $P^j{}_k = \delta^j{}_k - n^j n_k$ is the projection tensor orthogonal to n^j, *i.e.* show that for any vector V^j, (i) $P^j{}_k V^k$ is orthogonal to n^j, and (ii) $P^j{}_k P^k{}_l V^l = P^j{}_k V^k$.

(b) Show that the TT gauge $\bar{h}^{\mathrm{TT}}_{ij}$ of Eqs. (9.84)–(9.86) is related to the original $\bar{h}_{kl}$ of Eq. (9.83) by

$$\bar{h}^{\mathrm{TT}}_{ij} = P^k{}_i P^l{}_j \bar{h}_{kl} - \tfrac{1}{2} P_{ij}(P^{kl}\bar{h}_{kl}), \tag{9.163}$$

where n^j points in the z direction.

Show that $\text{\it ł}_{jk}$ is trace free, *i.e.* $\text{\it ł}^l{}_l = 0$.

For the systems described in Exer. 31, calculate the transverse-traceless quadrupole radiation field, Eqs. (9.85)–(9.86) or (9.163), along the x, y, and z axes. In Eqs. (9.85)–(9.86) be sure to change the indices appropriately when doing the calculation on the x and y axes, as in the discussion leading to Eq. (9.91).

Use Eq. (9.163) or a rotation of the axes in Eqs. (9.85)–(9.86) to calculate the amplitude and orientation of the polarization ellipse of the radiation from the simple oscillator, Eq. (9.88), traveling at an angle θ to the x axis.

The ω and 2ω terms in Eq. (9.93) are qualitatively different, in that the 2ω term depends only on the amplitude of oscillator A, while the ω term depends on both A and the separation of the masses l_0. Why should l_0 be involved – the masses don't move over that distance? The answer is that stresses *are* transmitted over that distance by the spring, and stresses cause the radiation. To see this, do an analogous calculation for a similar system, in which stresses are *not* passed over large distances. Consider a system consisting of two pairs of masses. Each pair has one particle of mass m and another of mass $M \gg m$. The masses within each pair are connected by a short spring whose natural frequency is ω. The pairs' centers of mass are at rest relative to one another. The springs oscillate with equal amplitude in such a way that each mass m oscillates sinusoidally with amplitude A, and the centers of oscillation of the masses are separated by $l_0 \gg A$. The masses oscillate out of phase. Use the calculation of Exer. 31(h) to show that the radiation field of the system is Eq. (9.93) without the ω term. The difference between this system and that in Eq. (9.93) may be thought to be the origin of the stresses to maintain the motion of the masses m.

Do the same as Exer. 36 for the binary system, Eqs. (9.98)–(9.99), but instead of finding the orientation of the linear polarization, find the orientation of the ellipse of elliptical polarization.

Let two spherical stars of mass m and M be in elliptical orbit about one another in the $x-y$ plane. Let the orbit be characterized by its total energy E and its angular momentum L.

(a) Use Newtonian gravity to calculate the equation of the orbits of both masses about their center of mass. Express the orbital period P, minimum separation l_0, and eccentricity e as functions of E and L.

(b) Calculate $\text{\sout{$I$}}_{kj}$ for this system.

(c) Calculate from Eq. (9.106) the TT radiation field along the x and z axes. Show that your result reduces to Eqs. (9.98)–(9.99) when $m = M$ and the orbits are circular.

Show from Eq. (9.101) that spherically symmetric motions produce no gravitational radiation.

Derive Eq. (9.115) from Eq. (9.114) in the manner suggested in the text.

(a) Derive Eq. (9.116).

(b) Derive Eqs. (9.117) and (9.118) by superposing Eqs. (9.107) and (9.116) and assuming R is small.

(c) Derive Eq. (9.120) in the indicated manner.

Show that if we define an averaged stress–energy tensor for the waves

$$T_{\alpha\beta} = \ll \bar{h}^{TT}_{\mu\nu,\alpha} \bar{h}^{TT\mu\nu}{}_{,\beta} \gg /32\pi \tag{9.164}$$

(where $\ll \gg$ denotes an average over both one period of oscillation in time and one wavelength of distance in all spatial directions), then the flux F of Eq. (9.122) is the component T^{0z} for that wave. A more detailed argument shows that Eq. (9.164) can in fact be regarded as the stress–energy tensor of any wave packet, provided the averages are defined suitably. This is called the Isaacson stress–energy tensor. See Misner *et al.* (1973) for details.

(a) Derive Eq. (9.125) from Eq. (9.123).

(b) Justify Eq. (9.127) from Eq. (9.125).

(c) Derive Eq. (9.127) from Eq. (9.122) using Exer. 33(b).

(a) Consider the integral in Eq. (9.128). We shall do it by the following method. (i) Argue on grounds of symmetry that $\int n^j n^k \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi$ must be proportional to δ^{jk}. (ii) Evaluate the constant of proportionality by explicitly doing the case $j = k = z$.

(b) Follow the same method for Eq. (9.129). In (i) argue that the integral can depend only on δ^{ij}, and show that the given tensor is the only one constructed purely from δ^{ij} that has the symmetry of being unchanged when the values of any two of its indices are exchanged.

Derive Eqs. (9.130) and (9.131), remembering Eq. (9.124) and the fact that $\rlap{-}I_{ij}$ symmetric.

(a) Recall that the angular momentum of a particle is p_ϕ. It follows that the angular momentum flux of a continuous system across a surface x^i = const. is $T_{i\phi}$. Use this and Exer. 43 to show that the total z component of angular momentum radiated by a source of gravitational waves (which is the integral over a sphere of large radius of $T_{r\phi}$ in Eq. (9.164)) is

$$F_J = -\tfrac{2}{5}(\ddot{I}_{xl}\dddot{I}_y{}^l - \ddot{I}_{yl}\dddot{I}_x{}^l). \tag{9.165}$$

(b) Show that if $\bar{h}^{TT}_{\mu\nu}$ depends on t and ϕ only as $\cos(\Omega t - m\phi)$, then the ratio of the total energy radiated to the total angular momentum radiated is Ω/m.

Calculate Eq. (9.135).

For the arbitrary binary system of Exer. 39:

(a) Show that the average energy loss rate over one orbit is

$$\langle \mathrm{d}E/\mathrm{d}t \rangle = -\frac{32}{5}\frac{\mu^2(m+M)^3}{a^5(1-e^2)^{7/2}}\left(1 + \frac{73}{24}e^2 + \frac{37}{96}e^4\right) \tag{9.166}$$

and from the result of Exer. 47(a)

$$\langle \mathrm{d}L/\mathrm{d}t \rangle = -\frac{32}{5}\frac{\mu^2(m+M)^{5/2}}{a^{7/2}(1-e^2)^{7/2}}\left(1+\frac{7}{8}e^2\right), \tag{9.167}$$

where $\mu = mM/(m+M)$ is the reduced mass.

(b) Show that

$$\langle \mathrm{d}a/\mathrm{d}t \rangle = -\frac{64}{5}\frac{\mu(m+M)^2}{a^3(1-e^2)^{7/2}}\left(1+\frac{73}{24}e^2+\frac{37}{96}e^4\right), \tag{9.168}$$

$$\langle \mathrm{d}e/\mathrm{d}t \rangle = -\frac{304}{15}\frac{\mu(m+M)^2 e}{a^4(1-e^2)^{5/2}}\left(1+\frac{121}{304}e^2\right), \tag{9.169}$$

$$\langle \mathrm{d}P/\mathrm{d}t \rangle = -\frac{192\pi}{5}\frac{\mu(m+M)^{3/2}}{a^{5/2}(1-e^2)^{7/2}}\left(1+\frac{73}{24}e^2+\frac{37}{96}e^4\right). \tag{9.170}$$

(c) Verify Eq. (9.144).

(Do parts (b) and (c) even if you can't do (a).) These were originally derived by Peters (1964).

1 In fact, the wavelength and frequency of light depend in any case on the observer, so the question cannot be posed in a frame-invariant way. This is another reason not to introduce them into discussions of how interferometers measure gravitational waves!

11

Schwarzschild geometry and black holes

11.1 Trajectories in the Schwarzschild spacetime

The ‘Schwarzschild geometry’ is the geometry of the vacuum spacetime outside a spherical star. It is determined by one parameter, the mass *M*, and has the line element

$$\mathrm{d}s^2 = -\left(1 - \frac{2M}{r}\right)\mathrm{d}t^2 + \left(1 - \frac{2M}{r}\right)^{-1}\mathrm{d}r^2 + r^2\,\mathrm{d}\Omega^2 \tag{11.1}$$

in the coordinate system developed in the previous chapter. Its importance is not just that it is the gravitational field of a star: we shall see that it is also the geometry of the spherical black hole. A careful study of its timelike and null geodesics – the paths of freely moving particles and photons – is the key to understanding the physical importance of this metric.

Black holes in Newtonian gravity

Before we embark on the study of fully relativistic black holes, it is well to understand that the physics is not really exotic, and that speculations on analogous objects go back two centuries. It follows from the weak equivalence principle, which was part of Newtonian gravity, that trajectories in the gravitational field of any body depend only on the position and velocity of the particle, not its internal composition. The question of whether a particle can escape from the gravitational field of a body is, then, only an issue of velocity: does it have the right escape velocity for the location it starts from. For a spherical body like a star, the escape velocity depends only on how far one is from the center of the body.

Now, a star is visible to us because light escapes from its surface. As long ago as the late 1700's, the British physicist John Michell and the French mathematician and physicist Pierre Laplace speculated (independently) on the possibility that stars might exist whose escape velocity was larger than the speed of light. At that period in history, it was popular to regard light as a particle traveling at a finite speed. Michell and Laplace both understood that if nature were able to make a star more compact than the Sun, but with the same mass, then it would have a larger escape velocity. It would therefore be possible in principle to make it compact enough for the escape velocity to be the velocity of light. The star would then be dark, invisible. For a spherical star, this is a simple computation. By conservation of energy, a particle launched from the surface of a star with mass M and radius R will just barely escape if its gravitational potential energy balances its kinetic energy (using Newtonian language):

$$\tfrac{1}{2}v^2 = \frac{GM}{R}. \tag{11.2}$$

Setting $v = c$ in this relation gives the criterion for the size of a star that would be invisible:

$$R = \frac{2GM}{c^2}. \tag{11.3}$$

Remarkably, as we shall see this is exactly the modern formula for the radius of a black hole in general relativity (§ 11.2). Now, both Michell and Laplace knew the mass of the Sun and the speed of light to enough accuracy to realize that this formula gives an absurdly small size, of order a few kilometers, so that to them the calculation was nothing more than an amusing speculation.

Today this is far more than an amusing speculation: objects of this size that trap light are being discovered all over the universe, with masses ranging from a few solar masses up to 10^9 or 10^{10} $M_\odot$. (We will discuss this in § 11.4 below.) The small size of a few kilometers is not as absurd as it once seemed. For a 1 $M_\odot$ star, using modern values for c and G, the radius is about 3 km. We saw in the last chapter that neutron stars have radii perhaps three times as large, with comparable masses, and that they cannot support more than three solar masses, perhaps less. So when neutron stars accrete large amounts of material, or when neutron stars merge together (as the stars in the Hulse–Taylor binary must do in about 10^8 y), formation of something even more compact is inevitable. What is more, it takes even less exotic physics to form a more massive black hole. Consider the mean density of an object (again in Newtonian terms) with the size given by Eq. (11.3):

$$\bar{\rho} = \frac{M}{\frac{4}{3}\pi R^3} = \frac{3c^6}{32\pi G^3 M^2}. \tag{11.4}$$

This scales as M^{-2}, so that the density needed to form such an object goes down as its mass goes up. It is not hard to show that an object with a mass of 10^9 $M_\odot$ would become a Newtonian 'dark star' when its density had risen only to the density of water! Astronomers believe that this is a typical mass for the black holes that are thought to power quasars (see § 11.4 below), so these objects would not necessarily require any exotic physics to form.

Although there is a basic similarity between the old concept of a Newtonian dark star and the modern black hole that we will explore in this chapter, there are big differences too. Most fundamentally, for Michell and Laplace the star was dark because light could not escape to infinity. The star was still there, shining light. The light would still leave the surface, but gravity would eventually pull it back, like a ball thrown upwards. In relativity, as we shall see, the light never leaves the 'surface' of a black hole; and this surface is itself not the edge of a massive body but just empty space, left behind by the inexorable collapse of the material that formed the hole.

Conserved quantities

We begin our study of relativistic black holes by examining the trajectories of particles. This will allow us eventually to see whether light rays are trapped or can escape.

We have seen (Eq. (7.29) and associated discussion) that when a spacetime has a certain symmetry, then there is an associated conserved momentum component for trajectories. Because our space has so many symmetries – time independence and spherical symmetry – the values of the conserved quantities turn out to determine the trajectory completely. We shall treat 'particles' with mass and 'photons' without mass in parallel.

Time independence of the metric means that the energy $-p_0$ is constant on the trajectory. For massive particles with rest mass $m \neq 0$, we define the energy per unit mass (specific energy) $\tilde{E}$, while for photons we use a similar notation just for the energy E:

$$\text{particle}: \; \tilde{E} := -p_0/m; \quad \text{photon}: E = -p_0. \tag{11.5}$$

Independence of the metric of the angle ϕ about the axis implies that the angular momentum p_ϕ is constant. We again define the specific angular momentum $\tilde{L}$ for massive particles and the ordinary angular momentum L for photons:

$$\text{particle } \tilde{L} := p_\phi/m; \quad \text{photon } L = p_\phi. \tag{11.6}$$

Because of spherical symmetry, motion is always confined to a single plane, and we can choose that plane to be the equatorial plane. Then θ is constant ($\theta = \pi/2$) for the orbit, so $\mathrm{d}\theta/\mathrm{d}\lambda = 0$, where λ is any parameter on the orbit. But p^θ is proportional to this, so it also vanishes. The other components of momentum are:

$$\begin{aligned} \text{particle}: p^0 &= g^{00} p_0 = m\left(1 - \frac{2M}{r}\right)^{-1} \tilde{E}, \\ p^r &= m\, \mathrm{d}r/\mathrm{d}\tau, \\ p^\phi &= g^{\phi\phi} p_\phi = m\frac{1}{r^2}\tilde{L}; \end{aligned} \tag{11.7}$$

$$\begin{aligned} \text{photon}: p^0 &= \left(1 - \frac{2M}{r}\right)^{-1} E, \\ p^r &= \mathrm{d}r/\mathrm{d}\lambda, \\ p^\phi &= \mathrm{d}\phi/\mathrm{d}\lambda = L/r^2. \end{aligned} \tag{11.8}$$

The equation for a photon's p^r should be regarded as *defining* the affine parameter λ. The equation $\vec{p} \cdot \vec{p} = -m^2$ implies particle :

$$-m^2\tilde{E}^2\left(1-\frac{2M}{r}\right)^{-1}+m^2\left(1-\frac{2M}{r}\right)^{-1}\left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right)^2 + \frac{m^2\tilde{L}^2}{r^2}=-m^2; \tag{11.9}$$

photon :

$$-E^2\left(1-\frac{2M}{r}\right)^{-1}+\left(1-\frac{2M}{r}\right)^{-1}\left(\frac{\mathrm{d}r}{\mathrm{d}\lambda}\right)^2+\frac{L^2}{r^2}=0. \tag{11.10}$$

These can be solved to give the basic equations for orbits,

$$\text{particle :}\quad \left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right)^2=\tilde{E}^2-\left(1-\frac{2M}{r}\right)\left(1+\frac{\tilde{L}^2}{r^2}\right); \tag{11.11}$$

$$\text{photon :}\quad \left(\frac{\mathrm{d}r}{\mathrm{d}\lambda}\right)^2=E^2-\left(1-\frac{2M}{r}\right)\frac{L^2}{r^2}. \tag{11.12}$$

Types of orbits

Both equations have the same general form, and we define the effective potentials

$$\text{particle}: \quad \tilde{V}^2(r) = \left(1 - \frac{2M}{r}\right)\left(1 + \frac{\tilde{L}^2}{r^2}\right); \tag{11.13}$$

$$\text{photon}: \quad V^2(r) = \left(1 - \frac{2M}{r}\right)\frac{L^2}{r^2}. \tag{11.14}$$

Their typical forms are plotted in Figs. 11.1 and 11.2, in which various points have been labeled and possible trajectories drawn (dotted lines).

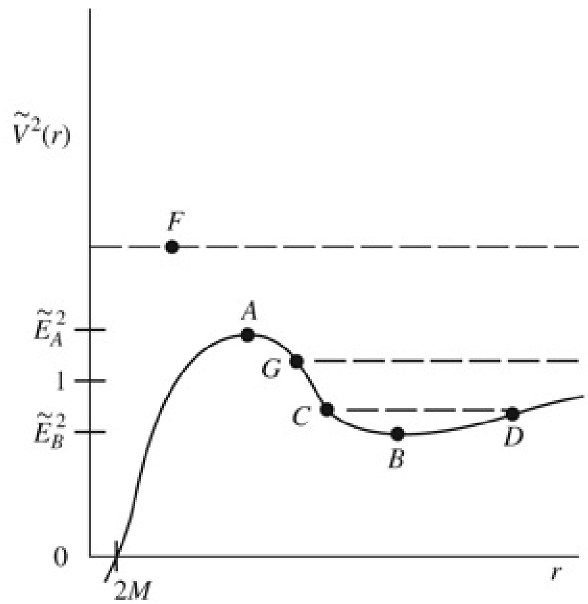


Figure 11.1 Typical effective potential for a massive particle of fixed specific angular momentum in the Schwarzschild metric.

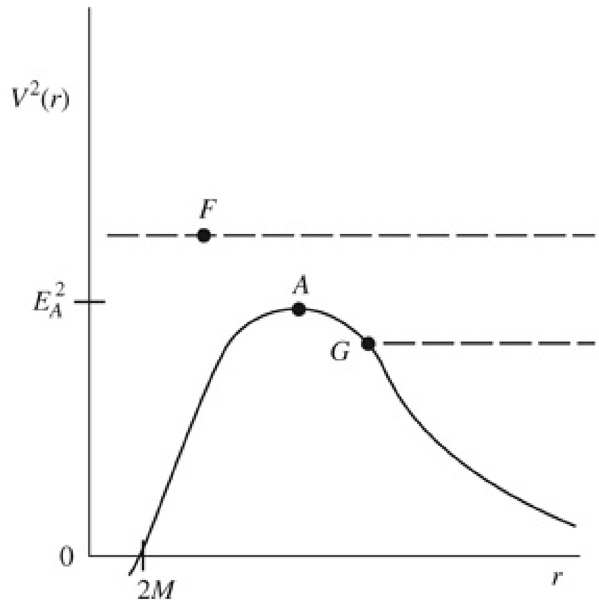


Figure 11.2 The same as Fig. 11.1 for a massless particle.

Both Eq. (11.11) and Eq. (11.12) imply that, since the left side is positive or zero, the energy of a trajectory must not be less than the potential V. (Here and until Eq. (11.17) we will take E and V to refer to $\tilde{E}$ and $\tilde{V}$ as well, since the remarks for the two cases are identical.) So for an orbit of given E, the radial range is restricted to those radii for which V is smaller than E. For instance, consider the trajectory which has the value of E indicated by point G (in either diagram). If it comes in from $r = \infty$, then it cannot reach smaller r than where the dotted line hits the V^2 curve, at point G. Point G is called a *turning* point. At G, since $E^2 = V^2$ we must have $(\mathrm{d}r/\mathrm{d}\lambda)^2 = 0$, from Eq. (11.12). Similar conclusions apply to Eq. (11.11). To see what happens here we differentiate Eqs. (11.11) and (11.12). For particles, differentiating the equation $\left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right)^2 = \tilde{E}^2 - \tilde{V}^2(r)$ with respect to τ gives

$$2\left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right)\left(\frac{\mathrm{d}^2r}{\mathrm{d}\tau^2}\right) = -\frac{\mathrm{d}\tilde{V}^2(r)}{\mathrm{d}r}\frac{\mathrm{d}r}{\mathrm{d}\tau},$$

or

$$\text{particles}: \qquad \frac{d^2 r}{d\tau^2} = -\frac{1}{2}\frac{d}{dr}\tilde{V}^2(r). \tag{11.15}$$

Similarly, the photon equation gives

$$\text{photons}: \qquad \frac{d^2 r}{d\lambda^2} = -\frac{1}{2}\frac{d}{dr}V^2(r). \tag{11.16}$$

These are the analogs in relativity of the equation

$$ma = -\nabla\phi,$$

where ϕ is the potential for some force. If we now look again at point G, we see that the radial acceleration of the trajectory is outwards, so that the particle (or photon) comes in to the minimum radius, but is accelerated outward as it turns around, and so it returns to $r = \infty$. This is a ‘hyperbolic’ orbit – the analog of the orbits which are true hyperbolae in Newtonian gravity.

It is clear from Eq. (11.15) or (Eq. (11.16) that a circular orbit (r = const.) is possible only at a minimum or maximum of V^2. These occur at points A and B in the diagrams (there is no point B for photons). A maximum is, however, unstable, since any small change in r results in an acceleration away from the maximum, by Eqs. (11.15) and (11.16). So for particles, there is one stable (B) and one unstable circular orbit (A) for this value of $\tilde{L}$. For photons, there is only one unstable orbit for this L. We can be quantitative by evaluating $0 = \frac{d}{dr}\left[\left(1 - \frac{2M}{r}\right)\left(1 + \frac{\tilde{L}^2}{r^2}\right)\right]$

and

$$0 = \frac{\mathrm{d}}{\mathrm{d}r}\left[\left(1 - \frac{2M}{r}\right)\frac{L^2}{r^2}\right].$$

These give, respectively

$$\text{particles}: \quad r = \frac{\tilde{L}^2}{2M}\left[1 \pm \left(1 - \frac{12M^2}{\tilde{L}^2}\right)^{1/2}\right]; \qquad (11.17)$$

$$\text{photons}: \quad r = 3M. \qquad (11.18)$$

For particles, there are two radii, as we expect, but only if $\tilde{L}^2 > 12M^2$. The two radii are identical for $\tilde{L}^2 = 12M^2$ and don't exist at all for $\tilde{L}^2 < 12M^2$. This indicates a qualitative change in the shape of the curve for $\tilde{V}^2(r)$ for small $\tilde{L}$. The two cases, $\tilde{L}^2 = 12M^2$ and $\tilde{L}^2 < 12M^2$, are illustrated in Fig. 11.3. Since there is a minimum $\tilde{L}^2$ for a circular particle orbit, there is also a minimum r, obtained by taking $\tilde{L}^2 = 12M^2$ in Eq. (11.17)

$$\text{particle}: \qquad r_{\mathrm{MIN}} = 6M. \qquad (11.19)$$

For photons, the unstable circular orbit is always at the same radius, $r = 3M$, regardless of L.

The last kind of orbit we need consider is the one whose energy is given by the line which passes through the point F in Figs. 11.1 and 11.2. Since this nowhere intersects the potential curve, this orbit plunges right through $r = 2M$ and never returns. From Exer. 1, § 11.7, we see that for such an orbit the impact parameter (b) is small: it is aimed more directly at the hole than are orbits of smaller $\tilde{E}$ and fixed $\tilde{L}$.

Of course, if the geometry under consideration is a star, its radius R will exceed $2M$, and the potential diagrams, Figs. 11.1–11.3, will be valid only outside R. If a particle reaches R, it will hit the star. Depending on R/M, then, only certain kinds of orbits will be possible.

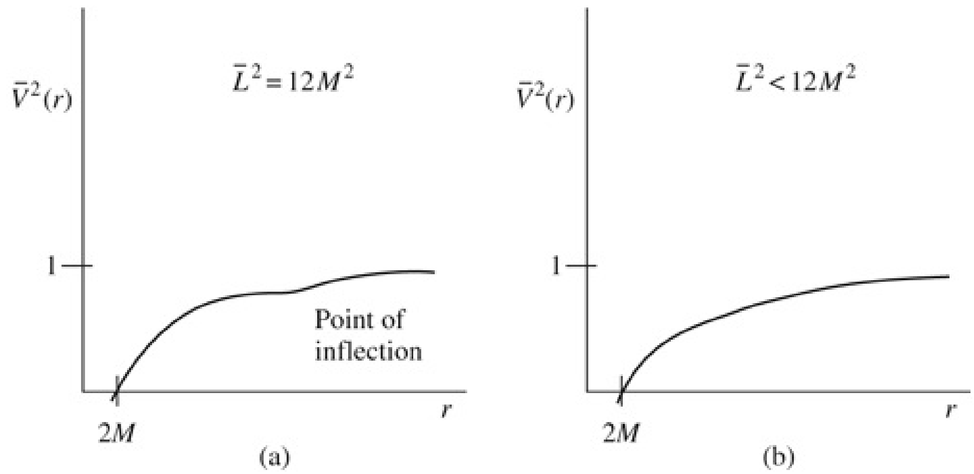


Figure 11.3 As Fig. 11.1 for the indicated values of specific angular momentum.

Perihelion shift

A particle (or planet) in a (stable) *circular* orbit around a star will make one complete orbit and come back to the same point (i.e. same value of ϕ) in a fixed amount of coordinate times, which is called its period P. This period can be determined as follows. From Eq. (11.17) it follows that a stable circular orbit at radius r has angular momentum

$$\tilde{L}^2 = \frac{Mr}{1 - 3M/r}, \tag{11.20}$$

and since $\tilde{E}^2 = \tilde{V}^2$ for a circular orbit, it also has energy

$$\tilde{E}^2 = \left(1 - \frac{2M}{r}\right)^2 \bigg/ \left(1 - \frac{3M}{r}\right). \tag{11.21}$$

Now, we have

$$\frac{\mathrm{d}\phi}{\mathrm{d}\tau} := U^\phi = \frac{p^\phi}{m} = g^{\phi\phi}\frac{p_\phi}{m} = g^{\phi\phi}\tilde{L} = \frac{1}{r^2}\tilde{L} \tag{11.22}$$

and

$$\frac{\mathrm{d}t}{\mathrm{d}\tau} := U^0 = \frac{p^0}{m} = g^{00}\frac{p_0}{m} = g^{00}(-\tilde{E}) = \frac{\tilde{E}}{1 - 2M/r}. \tag{11.23}$$

We obtain the angular velocity by dividing these:

$$\frac{\mathrm{d}t}{\mathrm{d}\phi} = \frac{\mathrm{d}t/\mathrm{d}\tau}{\mathrm{d}\phi/\mathrm{d}\tau} = \left(\frac{r^3}{M}\right)^{1/2}. \tag{11.24}$$

The period, which is the time taken for ϕ to change by 2π, is

$$P = 2\pi\left(\frac{r^3}{M}\right)^{1/2}. \tag{11.25}$$

This is the coordinate time, of course, not the particle's proper time. (But see Exer. 7, § 11.7: coordinate time is proper time far away.) It happens, coincidentally, that this is identical to the Newtonian expression.

Now, a slightly noncircular orbit will oscillate in and out about a central radius r. In Newtonian gravity the orbit is a perfect ellipse, which means, among other things, that it is *closed*: after a fixed amount of time it returns to the same point (same r and ϕ). In GR, this does not happen and a typical orbit is shown in Fig. 11.4. However, when the effects of relativity are small and the orbit is nearly circular, the relativistic orbit must be almost closed: it must look like an ellipse which slowly rotates about the center. One way to describe this is to look at the *perihelion* of the orbit, the point of closest approach to the star. ('Peri' means closest and 'helion' refers to the Sun; for orbits about any old star the name 'periastron' is more appropriate. For orbits around Earth – 'geo' – we speak of the 'perigee'. These opposite of 'peri' is 'ap': the furthest distance. Thus, an orbit also has an aphelion, apastron, or apogee, depending on what it is orbiting around. The general terms, not specific to any particular object, are perapsis and apapsis.) The perihelion will rotate around the star in some manner, and observers can hope to measure this. It has been measured for Mercury to be $43''$/century, and we must try to calculate it. Note that all other planets are further from the Sun and therefore under the influence of significantly smaller relativistic corrections to Newtonian gravity. The measurement of Mercury's precession is a herculean task, first accomplished in the 1800s. Due to various other effects, such as the perturbations of Mercury's orbit due to the other planets, the observed precession is about 5600″/century. The 43″ is only the part not explainable by Newtonian gravity, and Einstein's demonstration that his theory predicts exactly that amount was the first evidence in favor of the theory.

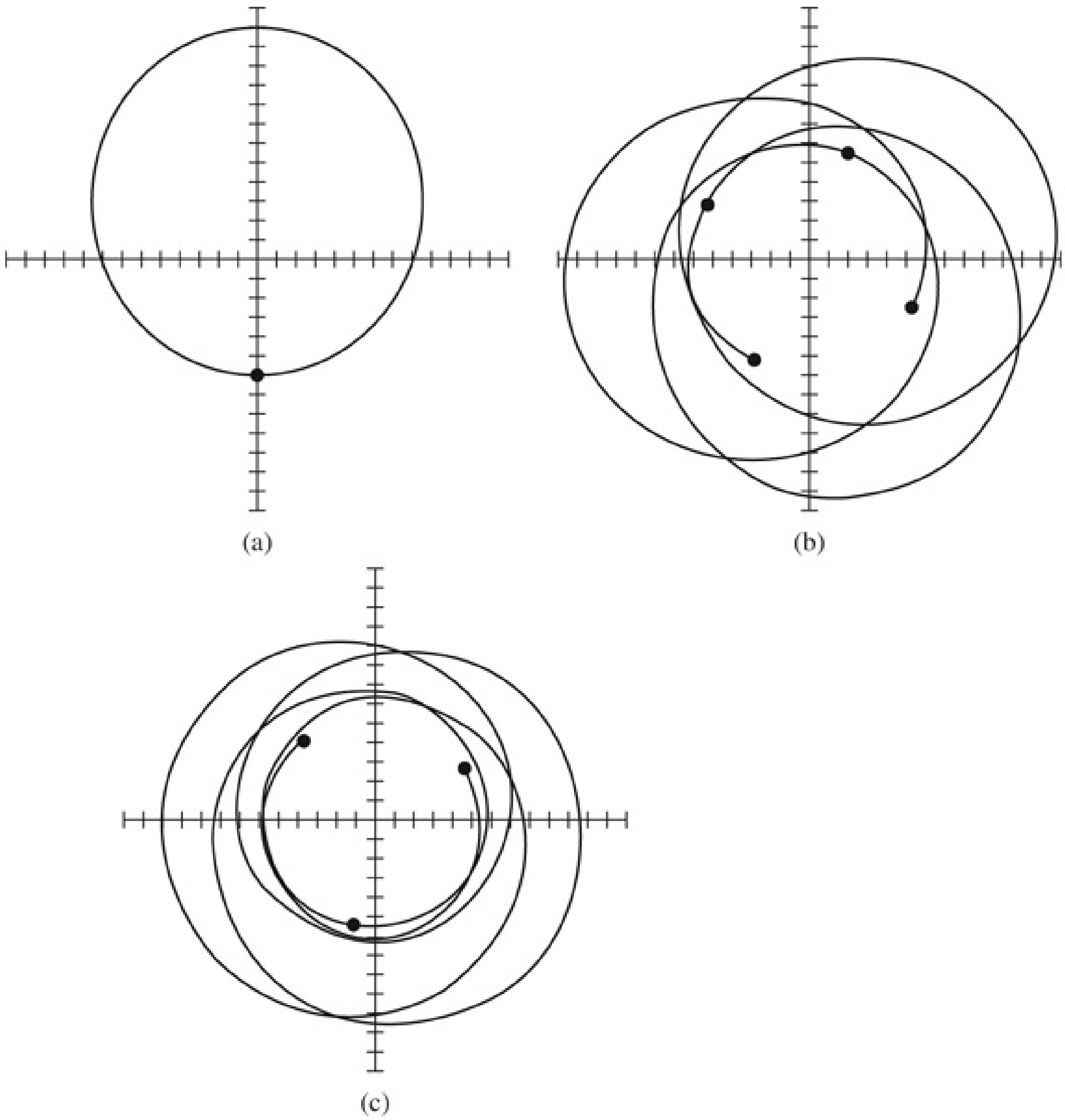


Figure 11.4 (a) A Newtonian orbit is a closed ellipse. Grid marked in units of *M*. (b) An orbit in the Schwarzschild metric with pericentric and apcentric distances similar to those in (a). Pericenters (heavy dots) advance by about 97° per orbit. (c) A moderately more compact orbit than in (b) has a considerably larger pericenter shift, about 130°.

To calculate the precession, let us begin by getting an equation for the particle's orbit. We have dr/dτ from Eq. (11.11). We get dϕ/dτ from Eq. (11.22) and divide to get

$$\left(\frac{\mathrm{d}r}{\mathrm{d}\phi}\right)^2 = \frac{\tilde{E}^2 - (1 - 2M/r)\left(1 + \tilde{L}^2/r^2\right)}{\tilde{L}^2/r^4}. \tag{11.26}$$

It is convenient to define

$$u := 1/r \tag{11.27}$$

and obtain

$$\left(\frac{\mathrm{d}u}{\mathrm{d}\phi}\right)^2 = \frac{\tilde{E}^2}{\tilde{L}^2} - (1 - 2Mu)\left(\frac{1}{\tilde{L}^2} + u^2\right). \tag{11.28}$$

The *Newtonian* orbit is found by neglecting u^3 terms (see Exer. 11, § 11.7)

$$\text{Newtonian}: \left(\frac{\mathrm{d}u}{\mathrm{d}\phi}\right)^2 = \frac{\tilde{E}^2}{\tilde{L}^2} - \frac{1}{\tilde{L}^2}(1 - 2Mu) - u^2. \tag{11.29}$$

A circular orbit in Newtonian theory has $u = M/\tilde{L}^2$ (take the square root equal to 1 in Eq. (11.17)), so we define

$$y = u - \frac{M}{\tilde{L}^2}, \tag{11.30}$$

so that y represents the deviation from circularity. We then get

$$\left(\frac{\mathrm{d}y}{\mathrm{d}\phi}\right)^2 = \frac{\tilde{E}^2 - 1}{\tilde{L}^2} + \frac{M^2}{\tilde{L}^4} - y^2. \tag{11.31}$$

It is easy to see that this is satisfied by

$$\text{Newtonian}: \quad y = \left[\frac{\tilde{E}^2 + M^2/\tilde{L}^2 - 1}{\tilde{L}^2}\right]^{1/2} \cos(\phi + B), \tag{11.32}$$

where B is arbitrary. This is clearly periodic: as ϕ advances by 2π, y returns to its value and, therefore, so does r. The constant B just determines the initial orientation of the orbit. It is interesting, but unimportant for our purposes, that by solving for r we get

$$\text{Newtonian}: \quad \frac{1}{r} = \frac{M}{\tilde{L}^2} + \left[\frac{\tilde{E}^2 + M^2/\tilde{L}^2 - 1}{\tilde{L}^2}\right]^{1/2} \cos(\phi + B), \tag{11.33}$$

which is the equation of an *ellipse*.

We now consider the relativistic case and make the same definition of y, but instead of throwing away the u^3 term in Eq. (11.28) we assume that the orbit is nearly circular, so that y is small, and we neglect only the terms in y^3. Then we get Nearly circular:

$$\left(\frac{\mathrm{d}y}{\mathrm{d}\phi}\right)^2 = \frac{\tilde{E}^2 + M^2/\tilde{L}^2 - 1}{\tilde{L}^2} + \frac{2M^4}{\tilde{L}^6} + \frac{6M^3}{\tilde{L}^2}y - \left(1 - \frac{6M^2}{\tilde{L}^2}\right)y^2. \tag{11.34}$$

This can be made analogous to Eq. (11.31) by completing the square on the right-hand side. The result is the solution

$$y = y_0 + A\cos(k\phi + B), \tag{11.35}$$

where B is arbitrary and the other constants are

$$\begin{aligned} k &= \left(1 - \frac{6M^2}{\tilde{L}^2}\right)^{1/2}, \\ y_0 &= 3M^3/k^2\tilde{L}^2, \\ A &= \frac{1}{k}\left[\frac{\tilde{E}^2 + M^2/\tilde{L}^2 - 1}{\tilde{L}^2} + \frac{2M^4}{\tilde{L}^6} - y_0^2\right]^{1/2}. \end{aligned} \tag{11.36}$$

The appearance of the constant y_0 just means that the orbit oscillates not about $y = 0$ ($u = \tilde{M}/L^2$) but about $y = y_0$: Eq. (11.30) doesn't use the correct radius for a circular orbit in GR. The amplitude A is also somewhat different, but what is most interesting here is the fact that k is not 1. The orbit returns to the same r when $k\phi$ goes through 2π, from Eq. (11.35). Therefore the change in ϕ from one perihelion to the next is

$$\Delta\phi = \frac{2\pi}{k} = 2\pi\left(1 - \frac{6M^2}{\tilde{L}^2}\right)^{-1/2}, \tag{11.37}$$

which, for nearly Newtonian orbits, is

$$\Delta\phi \simeq 2\pi\left(1 + \frac{3M^2}{\tilde{L}^2}\right). \tag{11.38}$$

The perihelion *advance*, then, from one orbit to the next, is

$$\Delta\phi = 6\pi M^2/\tilde{L}^2 \text{ radians per orbit.} \tag{11.39}$$

We can use Eq. (11.20) to obtain $\tilde{L}$ in terms of r, since the corrections for noncircularity will make changes in Eq. (11.39) of the same order as terms we have already neglected. Moreover, if we consider orbits about a nonrelativistic star, we can approximate Eq. (11.20) by $\tilde{L}^2 = \dfrac{Mr}{1 - 3M/r} \approx Mr,$

so that we get

$$\Delta\phi \approx 6\pi \frac{M}{r}. \tag{11.40}$$

For Mercury's orbit, $r = 5.55 \times 10^7$ km and $M = 1\, M_\odot = 1.47$ km, so that

$$(\Delta\phi)_{\text{Mercury}} = 4.99 \times 10^{-7} \text{ radians per orbit.} \tag{11.41}$$

Each orbit take 0.24 yr, so the shift is

$$(\Delta\phi)_{\text{Mercury}} = .43''/\text{yr} = 43''/\text{century}. \tag{11.42}$$

Binary pulsars

Another system in which the pericenter shift is observable is the Hulse-Taylor binary pulsar system PSR B1913+16 that we introduced in § 9.3. While this is not a "test particle" orbiting a spherical star, but is rather two roughly equal-mass stars orbiting their common center of mass, the pericenter shift (in this case, the periastron shift) still happens in the same way. The two neutron stars of the Hulse-Taylor system have mean separation 1.2×10^9 m, so using Eq. (11.40) with $M = 1.4\, M_\odot = 2.07$ km gives a crude estimate of $\Delta\phi = 3.3 \times 10^{-5}$ radians per orbit = $2^\circ.1$ per year. This is much easier to measure than Mercury's shift! In fact, a more careful calculation, taking into account the high eccentricity of the orbit and the fact that the two stars are of comparable mass, predict 4°.2 per year.

For our purposes here we have calculated the periastron shift from the known masses of the star. But in fact the observed shift of 4.2261° ± 0.0007 per year is one of the data which enable us to calculate the masses of the neutron stars in the PSR B1913+16 system. The other datum is another relativistic effect: a redshift of the signal which results from two effects. One is the special-relativistic 'transverse-Doppler' term: the $0(v)^2$ term in Eq. (2.39). The other is the changing gravitational redshift as the pulsar's eccentric orbit brings it in and out of its companion's gravitational potential. These two effects are observationally indistinguishable from one another, but their combined resultant redshift gives one more number which depends on the masses of the stars. Using it and the periastron shift and the Newtonian mass function for the orbit allows us to determine the stars' masses and the orbit's inclination (see Stairs 2003).

While PSR B1913+16 was the first binary pulsar to be discovered, astronomers now know of many more (Lorimer 2008, Stairs 2003). The most dramatic system is the so-called double pulsar system PSR J0737-3039, the first binary discovered in which *both* members are seen as radio pulsars (Kramer *et al*. 2006). In this system the pericenter shift is around 17° per year! Because both pulsars can be tracked, this system promises to become the best testing ground so far for the deviations of general relativity from simple Newtonian behavior. We discuss some of these now.

Post-Newtonian gravity

The pericenter shift is an example of corrections that general relativity makes to Newtonian orbital dynamics. In going from Eq. (11.37) to Eq. (11.38), we made an approximation that the orbit was 'nearly Newtonian', and we did a Taylor expansion in the small quantity $M^2/\tilde{L}^2$, which by Eq. (11.20) is $(M/r)(1 - 3M/r)$. This is indeed small if the particle's orbit is far from the black hole, $r \gg M$. So orbits far from the hole are very nearly like Newtonian orbits, with small corrections like the pericenter shift. These are orbits where, in Newtonian language, the gravitational potential M/r is small, as is the particle's orbital velocity, which for a circular orbit is related to the potential by $v^2 = M/r$. Far from even an extremely relativistic source, gravity is close to being Newtonian. We say that the pericenter shift is a post-Newtonian effect, a correction to Newtonian motion in the limit of weak fields and slow motion.

We showed in § (8.4) that Newton's field equations emerge from the full equations of general relativity in this same limit, where the field (we called it *h* in that calculation) is weak and the velocities small. It is possible to follow that approximation to higher order, to keep the first post-Newtonian corrections to the Einstein field equations in this same limit. If we did this, we would be able to show the result we have asserted, that the pericenter shift is a general feature of orbital motion. We showed in that section that the metric of spacetime that describes a Newtonian system with gravitational potential ϕ is, to first order in ϕ, given by the metric

$$ds^2 = -(1 + 2\phi)dt^2 + (1 - 2\phi)(dx^2 + dy^2 + dz^2). \tag{11.43}$$

For comparison, let us take the limit for weak fields of the Schwarzschild metric given in Eq. (11.1:

$$ds^2 \approx -\left(1 - \frac{2M}{r}\right) dt^2 + \left(1 + \frac{2M}{r}\right) dr^2 + r^2\, d\Omega^2,$$

$$ \tag{11.44} $$

where we have expanded g_{rr} and kept only the lowest order term in M/r. These two metrics look similar – their g_{00} terms are identical – but they appear not to be identical, because if we transform the spatial line element of Eq. (11.43) to polar coordinates in the usual way, we would expect to see in Eq. (11.44) the spatial line element

$$ \mathrm{d}\ell^2 = \left(1 + \frac{2M}{r}\right)\left(\mathrm{d}r^2 + r^2\,\mathrm{d}\Omega^2\right). \tag{11.45} $$

(Readers who want to be reminded about the transformation from Cartesian to spherical coordinates in flat space should see Eq. (6.19) and Exer. 28, § 6.9.) Is this difference from the spatial part of Eq. (11.44) an indication that the Schwarzschild metric does not obey the Newtonian limit? The answer is no – we are dealing with a coordinate effect.

Coordinates in which the line element has the form of Eq. (11.45) are called *isotropic coordinates* because there is no distinction in the metric among the directions x, y, and z. Isotropic coordinates are related to Schwarzschild coordinates by a change in the definition of the radial coordinate. If we call the isotropic coordinates $(t,\ \bar{r},\ \theta,\ \phi)$, where $(t,\ \theta,\ \phi)$ are the same as in Schwarzschild, then we make the simple transformation for large r given by

$$ \bar{r} = r - M. \tag{11.46} $$

Then to lowest order in M/r or $M/\bar{r}$, the expression $1 + 2M/r$ in Eq. (11.44) equals $1 - 2M/\bar{r}$. But the factor in front of $\mathrm{d}\Omega^2$ is r^2, which becomes (again to first order) $\bar{r}^2(1 + 2M/r)$. It follows that this simple transformation changes the spatial part of the line element of Eq. (11.44) to Eq. (11.45) in terms of the radial coordinate $\bar{r}$. This demonstrates that the far field of the Schwarzschild solution does indeed conform to the form in Eq. (8.50).

There are of course many other post-Newtonian effects in general relativity besides the pericenter shift. In the next paragraph we will explore the deflection of light. Later in this chapter we will meet the dragging of inertial frames due to the rotation of the source of gravity (also called gravitomagnetism). It is possible to expand the Einstein equations beyond the first post-Newtonian equations and find second and higher post-Newtonian effects. Physicists have put considerable work into doing such expansions, because they provide highly accurate predictions of the orbits of inspiralling neutron stars and black holes, which are needed for the detection of the gravitational waves they emit (recall Ch. 9).

The post-Newtonian expansion assumes not just weak gravitational fields but also slow velocities. The effect of this is that the Newtonian (Keplerian) motion of a planet depends only on the g_{00} part of the metric. The reason is that, in computing the total elapsed proper time along the world line of the particle, the spatial distance increments the particle makes (for example $\mathrm{d}r$) are much smaller than the time increments $\mathrm{d}t$, since $\mathrm{d}r/\mathrm{d}t \ll 1$. Recall that g_{00} is also responsible for the gravitational redshift. It follows, therefore, that Newtonian gravity can be identified with the gravitational redshift: knowing one determines the other fully. *Newtonian gravity is produced exclusively by the curvature of time in spacetime.* Spatial curvature comes in only at the level of post-Newtonian corrections.

Gravitational deflection of light

In our discussion of orbits we treated only particles, not photons, because photons do not have bound orbits in Newtonian gravity. In this section we treat the analogous effect for photons, their deflection from straight-line motion as they pass through a gravitational field. Historically, this was the first general-relativistic effect to have been predicted before it was observed, and its confirmation in the eclipse of 1919 (see McCrae 1979) made Einstein an international celebrity. The fact that it was a British team (led by Eddington) who made the observations to confirm the theories of a German physicist incidentally helped to alleviate post-war tension between the scientific communities of the two countries. In modern times, the light-deflection phenomenon has become a key tool of astronomy, as we describe in a separate paragraph on gravitational lensing below. But first we need to understand how and why gravity deflects light.

We begin by calculating the trajectory of a photon in the Schwarzschild metric under the assumption that M/r is everywhere small along the trajectory. The equation of the orbit is the ratio of Eq. (11.8) to the square root of Eq. (11.12):

$$\frac{\mathrm{d}\phi}{\mathrm{d}r} = \pm\frac{1}{r^2}\left[\frac{1}{b^2} - \frac{1}{r^2}\left(1 - \frac{2M}{r}\right)\right]^{-1/2}, \tag{11.47}$$

where we have defined the *impact parameter*,

$$b := L/E. \tag{11.48}$$

In Exer. 1, § 11.7, it is shown that b would be the minimum value of r in Newtonian theory, where there is no deflection. It therefore represents the 'offset' of the photon's initial trajectory from a parallel one moving purely radially. An incoming photon with $L > 0$ obeys the equation

$$\frac{\mathrm{d}\phi}{\mathrm{d}u} = \left(\frac{1}{b^2} - u^2 + 2Mu^3\right)^{-1/2}, \tag{11.49}$$

with the same definition as before,

$$u = 1/r. \tag{11.50}$$

If we neglect the u^3 term in Eq. (11.49), all effects of M disappear, and the solution is

$$r \sin(\phi - \phi_0) = b, \tag{11.51}$$

a straight line. This is, of course, the Newtonian result.

Suppose now we assume $Mu \ll 1$ but not entirely negligible. Then if we define

$$y := u(1 - Mu), \qquad u = y(1 + My) + 0(M^2u^2), \tag{11.52}$$

Eq. (11.49) becomes

$$\frac{d\phi}{dy} = \frac{(1 + 2My)}{\left(b^{-2} - y^2\right)^{1/2}} + 0(M^2u^2). \tag{11.53}$$

This can be integrated to give

$$\phi = \phi_0 + \frac{2M}{b} + \arcsin(by) - 2M\left(\frac{1}{b^2} - y^2\right)^{1/2}. \tag{11.54}$$

The initial trajectory has $y \to 0$, so $\phi \to \phi_0$: ϕ_0 is the incoming direction. The photon reaches its smallest r when $y = 1/b$, as we can see from setting $dr/d\lambda = 0$ in Eq. (11.22) and using our approximation $Mu \ll 1$. This occurs at the angle $\phi = \phi_0 + 2M/b + \pi/2$. It has thus passed through an angle $\pi/2 + 2M/b$ as it travels to its point of closest approach. By symmetry, it passes through a further angle of the same size as it moves outwards from its point of closest approach (see Fig. 11.5). It thus passes through a total angle of $\pi + 4M/b$. If it were going on a straight line, this angle would be π, so the net deflection is

$$\Delta\phi = 4M/b. \tag{11.55}$$

To the accuracy of our approximations, we may use for b the radius of closest approach rather than the impact parameter L/E. For the sum, the maximum effect is for trajectories for which $b = R_\odot$, the radius of the Sun. Given $M = 1\ M_\odot = 1.47$ km and $R_\odot = 6.96 \times 10^5$ km, we find

$$(\Delta\phi)_{\odot,\max} = 8.45 \times 10^{-6}\ \text{rad} = 1''.74. \tag{11.56}$$

For Jupiter, with $M = 1.12 \times 10^{-3}$ km and $R = 6.98 \times 10^4$ km, we have

$$(\Delta\phi)_{\Psi,\max} = 6.42 \times 10^{-8}\ \text{rad} = 0''.013. \tag{11.57}$$

This deflection has been measured by the Hipparcos astrometry satellite.

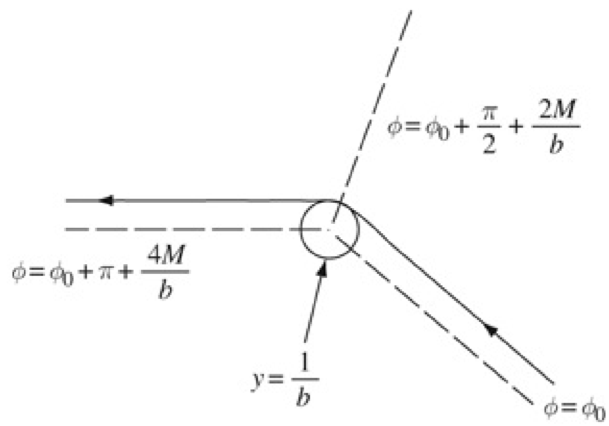


Figure 11.5 Deflection of a photon.

Of course, satellite observations of stellar positions are made from a position near Earth, and for stars that are not near the Sun in the sky the satellite will receive their light before the total deflection, given by Eq. (11.55), has taken place. This situation is illustrated in Fig. 11.6. An observer at *rest* at the position of the satellite observes an apparent position in the direction of the vector $\vec{a}$, tangent to the path of the light ray, and if he knows his distance r from the Sun, he can calculate the true direction to the star, $\vec{s}$. Exer. 16, § 11.7, derives the general result.

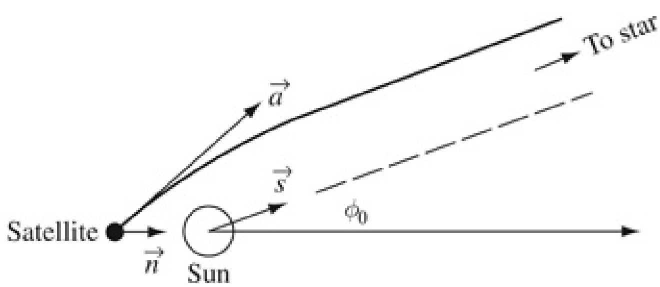


Figure 11.6 An observation from Earth of a star not at the limb of the Sun does not need to correct for the full deflection of Fig. 11.5.

Although the deflection of light was a key to establishing the correctness of general relativity, it is interesting that there is a purely Newtonian argument for light deflection. This was first predicted by Cavendish in 1784 and independently by the German astronomer J. G. von Söldner (1776–1833) in 1801. The argument relies on light behaving like a particle moving at speed c. Since the motion of a particle in a gravitational field depends only on its velocity, Cavendish and von Söldner were able to compute the simple result that the deflection would be $2M/b$, exactly *half* of the prediction of general relativity in Eq. (11.55). Einstein himself, unaware of this previous work, derived the same result in 1908, just at the beginning of his quest for a relativistic theory of gravity. So the triumph of general relativity was not that it predicted a deflection, but that it predicted the right amount of deflection.

The reason that this Newtonian result is half of the fully relativistic one is not hard to understand. Recall the remark at the end of the paragraph on post-Newtonian effects, that the orbits of planets depend only on g_{00} because their velocities are small. Now, light is not slow: the spatial increments dr are comparable to the time increments dt. This means that the spatial part of the metric, say g_{rr}, are of equal weight with g_{00} in determining the motion of a photon. Looking at Eq. (11.44), we see that the deviations from flatness in both g_{00} and g_{rr} are the same. Whatever deflection is produced by g_{00} (the only part of the metric that the Newtonian computations were sensitive to) is doubled by g_{rr}. *The extralarge deflection of light in general relativity compared with Newtonian gravity is direct evidence for the curvature of space as well as time.*

Gravitational lensing

It may of course happen that photons from the same star will travel trajectories that pass on opposite sides of the deflecting star and intersect each other after deflection, as illustrated in Fig. 11.7. Rays 1 and 2 are essentially parallel if the star ($*$) is far from the deflecting object (S). An observer at position B would then see *two* images of the star, coming from apparently different directions.

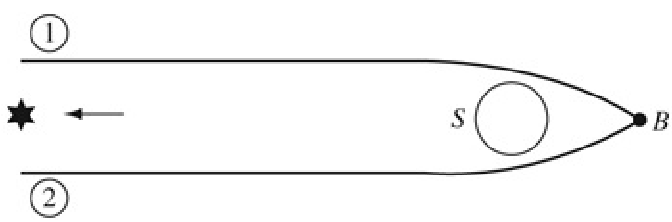


Figure 11.7 Deflection can produce multiple images.

This is a very simple and special arrangement of the objects, but it illustrates the principle that gravitating bodies act as lenses. Lensing is essentially universal: no matter how weak the deflection, it would always be possible to place the observer B far enough away from S to see two rays from the same point on the source. We don't get such "double vision" when we look at the heavens because the probability of 'being in exactly the right spot' B for any given star and lens is small, and because many sources are not pointlike like our star: if the angular separation of the images at *B* (which is of order the deflection angle) is small compared to the angular size of the object on the sky, then we are not likely to be able to tell the difference between the two rays.

But as astronomers have built larger and more powerful telescopes, able to see much greater distances into the universe, they have revealed a sky filled with lensed images. Of particular importance is lensing by clusters of galaxies. There are so many galaxies in the universe that, beyond any given relatively distant cluster, there is a high probability that there will be another group of galaxies located in just the right position to be lensed into multiple images in our telescope: the probability of 'being in exactly the right spot' has become reasonably large. What is more, the masses of galaxy clusters are huge, so the deflections are much bigger than the sizes of the more distant galaxy images, so separating them is not a problem. In fact, we often see multiple images of the same object, created by the irregularities of the lensing mass distribution. A good example is in Fig. 11.8.

Figure 11.8 A picture taken with the Hubble Space Telescope of a cluster called 0024+1654 (fuzzy round galaxies) showing many images of a much more distant galaxy. The images are distorted, all of them stretched in the direction transverse to the line joining the image and the center of the cluster, and compressed along this line. Picture courtesy of W. N. Colley and E. Turner (Princeton University), J. A. Tyson (Bell Labs, Lucent Technologies) and NASA.

Even more important than the creation of separate images can be the brightening of single images by the focusing of light from them. This is called magnification. The magnification of galaxy images by lensing makes it possible for astronomers to see galaxies at greater distances, and has helped studies of the very early universe. Lensing also helps us map the mass distribution of the lensing cluster, and this has shown that clusters have much more mass than can be associated with their luminous stars. Astronomers call this *dark matter*, and it will be an important subject in the next chapter, when we discuss cosmology.

Detailed modeling of observed lenses has shown that the dark matter is distributed more smoothly inside clusters than the stars, which clump into the individual galaxies. The mass of a cluster may be ten or more times the mass of its stars. The dark matter is presumably made of something that carries no electric charge, because light from the distant galaxies passes through it without absorption or scattering, and it seems to emit no electromagnetic radiation. Physicists do not know of any elementary particle that could serve as dark matter in this way: apart from neutrinos, all electrically neutral particles known at present (2008) are unstable and decay quickly. And neutrinos are thought to be too light to have been trapped in the gravitational fields of clusters. Dark matter is one of the biggest mysteries at the intersection of astronomy and particle physics today.

Studies of gravitational lensing have the potential to reveal any matter that clumps in some way. Lensing observations of individual stars in our galaxy (microlensing) have shown that there appears to be a population of objects with stellar masses that are also dark: they are observed only because they pass across more distant stars in our Galaxy and produce lensing. In this case, observers do not see two separate images. Instead, they see is brightening of the distant star, caused by the fact that the lens focuses more light on the telescope than would reach it if the lens were not there. This brightening is temporary because the lensing object itself moves, and so the lensing event is transitory. Optical studies of the region around the lensed stars have shown that the lensing object is not as bright as a normal white dwarf star but has a mass that is a good fraction of a solar mass. The nature of these lensing objects, which could represent a significant fraction of the mass of the Galaxy, is still (2008) a mystery.

11.2 Nature of the surface $r = 2M$

Coordinate singularities

It is clear that something funny goes wrong with the line element, Eq. (11.1) at $r = 2M$, but what is not clear is whether the problem is with the geometry or just with the coordinates. Coordinate singularities – places where the coordinates don't describe the geometry properly – are not unknown in ordinary calculus. Consider spherical coordinates at the poles. The north pole on a sphere has coordinates $\theta = 0, 0 \leqslant \phi < 2\pi$. That is, although ϕ can have any value for $\theta = 0$, all values really correspond to a single point. We might draw a coordinate diagram of the sphere as follows (Fig. 11.9 – maps of the globe are sometimes drawn this way), in which it would not be at all obvious that all points at $\theta = 0$ are really the same point. We *could*, however, convince ourselves of this by calculating the circumference of every circle of constant θ and verifying that these approached zero as $\theta \to 0$ and $\theta \to \pi$. That is, by asking questions that have an invariant geometrical meaning, we can tell if the coordinates are bad. For the sphere, the metric is positive-definite, so if two points have zero distance between them, they are the same point (e.g. $\theta = 0$, $\phi = \pi$ and $\theta = 0$, $\phi = 2\pi$: see Exer. 18, § 11.7). In relativity, the situation is more subtle, since there are curves (null curves) where distinct points have zero invariant distance between them. In fact, the whole question of the nature of the surface $r = 2M$ is so subtle that it was not answered satisfactorily until 1960. (This was just in time, too, since black holes began to be of importance in astronomy within a decade as new technology made observations of quasars, pulsars, and X-ray sources possible.) We shall explore the problem by asking a few geometrical questions about the metric and then demonstrating a coordinate system which has no singularity at this surface.

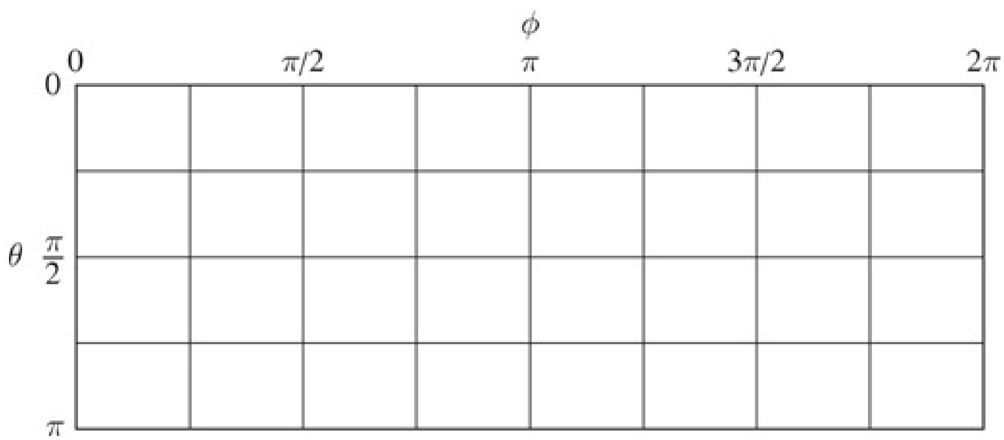


Figure 11.9 One way of drawing as sphere on a flat piece of paper. Not only are $\boldsymbol{\phi} = 0$ and $\boldsymbol{\phi} = 2\boldsymbol{\pi}$ really the same lines, but the lines $\boldsymbol{\theta} = 0$ and $\boldsymbol{\theta} = \boldsymbol{\pi}$ are each really just one point. Spherical coordinates are therefore not faithful representations of the sphere everywhere.

Infalling particles

Let a particle fall to the surface $r = 2M$ from any finite radius R. How much proper time does that take? That is, how much time has elapsed on the particle's clock? The simplest particle to discuss is the one that falls in radially. Since $d\phi = 0$, we have $\tilde{L} = 0$ and, from Eq. (11.11),

$$\left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right)^2 = \tilde{E}^2 - 1 + \frac{2M}{r}, \tag{11.58}$$

or

$$d\tau = -\frac{dr}{\left(\tilde{E}^2 - 1 + 2M/r\right)^{1/2}} \tag{11.59}$$

(the minus sign because the particle falls inward). It is clear that if $\tilde{E}^2 > 1$ (unbound particle), the integral of the right-hand side from R to $2M$ is finite. If $\tilde{E} = 1$ (particle falling from rest at ∞), the integral is simply

$$\Delta\tau = \frac{4M}{3}\left[\left(\frac{r}{2M}\right)^{3/2}\right]_{2M}^{R}, \tag{11.60}$$

which is again finite. And if $\tilde{E} < 1$, there is again no problem since the particle cannot be at larger r than where $1 - \tilde{E}^2 = 2M/r$ (see Eq. (11.58)). So the answer is that any particle can reach the horizon in a finite amount of proper time. In fact there is nothing about the integral that prevents us placing the lower limit smaller than $2M$, that is the other side of the surface $r = 2M$. The particle apparently can go inside $r = 2M$ in a finite proper time.

We now ask how much coordinate time elapses as the particle falls in. For this we use

$$U^0 = \frac{dt}{d\tau} = g^{00}U_0 = g^{00}\frac{p_o}{m} = -g^{00}\tilde{E} = \left(1 - \frac{2M}{r}\right)^{-1}\tilde{E}.$$

Therefore we have

$$\mathrm{d}t = \frac{\tilde{E}\,\mathrm{d}\tau}{1 - 2M/r} = -\frac{\tilde{E}\,\mathrm{d}r}{(1 - 2M/r)(\tilde{E}^2 - 1 + 2M/r)^{1/2}}. \tag{11.61}$$

For simplicity, we again consider the case $\tilde{E} = 1$ and examine this near $r = 2M$ by defining the new variable $\varepsilon := r - 2M$.

Then we get

$$\mathrm{d}t = \frac{-(\varepsilon + 2M)^{3/2}\mathrm{d}\varepsilon}{(2M)^{1/2}\varepsilon}. \tag{11.62}$$

It is clear that as $\varepsilon \to 0$ the integral of this goes like In ε, which diverges. We would also find this for $\tilde{E} \neq 1$, because the divergence comes from the $[1 - (2M/r)]^{-1}$ term, which doesn't contain $\tilde{E}$. Therefore a particle reaches the surface $r = 2M$ only after an infinite coordinate time has elapsed. Since the proper time is finite, the coordinate time must be behaving badly.

Inside $r = 2M$

To see just how badly it behaves, let us ask what happens to a particle after it reaches $r = 2M$. It must clearly pass to smaller r unless it is destroyed. This might happen if at $r = 2M$ there were a 'curvature singularity', where the gravitational forces grew strong enough to tear anything apart. But calculation of the components $R^{\alpha}{}_{\beta\mu\nu}$ of Riemannian tensor in the local inertial frame of the infalling particle shows them to be perfectly finite: Exer. 20, § 11.7. So we must conclude that the particle will just keep going. If we look at the geometry inside but near $r = 2M$, by introducing $\varepsilon := 2M - r$, then the line element is

$$\mathrm{d}s^2 = \frac{\varepsilon}{2M - \varepsilon}\mathrm{d}t^2 - \frac{2M - \varepsilon}{\varepsilon}\mathrm{d}\varepsilon^2 + (2M - \varepsilon)^2\mathrm{d}\Omega^2. \tag{11.63}$$

Since $\varepsilon > 0$ inside $r = 2M$, we see that a line on which t, θ, ϕ are constant has $\mathrm{d}s^2 < 0$: it is timelike. Therefore ε (and hence r) is a *timelike* coordnate, while t has become spacelike: even more evidence for the funniness of t and r! Since the infalling particle must follow a timelike world line, it must constantly change r, and of course this means *decrease* r. So a particle inside $r = 2M$ will inevitably reach $r = 0$, and there a *true* curvature singulaity awaits it: sure destruction by infinite forces (Exer. 20, § 11.7). But what happens if the particle inside $r = 2M$ tries to send out a photon to someone outside $r = 2M$ in order to describe his impending doom? This photon, no matter how directed, must also go forward in 'time' as seen locally by the particle, and this means to decreasing r. So the photon will not get out either. Everything inside $r = 2M$ is trapped and, moreover, doomed to encounter the singularity at $r = 0$, since $r = 0$ is in the future of every timelike and null world line inside $r = 2M$. Once a particle crosses the surface $r = 2M$, it cannot be seen by an external observer, since to be seen means to send out a photon which reaches the external observer. This surface is therefore called a *horizon*, since a horizon on Earth has the same effect (for different reasons!). We shall henceforth refer to $r = 2M$ as the Schwarzschild horizon.

Coordinate systems

So far, our approach has been purely algebraic – we have no ‘picture’ of the geometry. To develop a picture we will first draw a coordinate diagram in Schwarzschild coordinates, and on it we will draw the light cones, or at least the paths of the radially ingoing and outgoing null lines emanating from certain events (Fig. 11.10). These light cones may be calculated by solving $ds^2 = 0$ for θ and ϕ constant:

$$\frac{dt}{dr} = \pm\left(1 - \frac{2M}{r}\right)^{-1}. \tag{11.64}$$

In a $t - r$ diagram, these lines have slope ± 1 far from the star (standard SR light cone) but their slope approaches $\pm\infty$ as $r \to 2M$. This means that they become more vertical: the cone ‘closes up’. Since particle world lines are confined within the local light cone (a particle must move slower than light), this closing up of the cones forces the world lines of particles to become more vertical: if they reach $r = 2M$, they reach it at $t = \infty$. This is the ‘picture’ behind the algebraic result that a particle takes infinite coordinate time to reach the horizon. Notice that *no* particle world line reaches the line $r = 2M$ for any finite value of t. This might suggest that the line ($r = 2M$, $-\infty < t < \infty$) is really not a line at all but a single point in spacetime. That is, our coordinates may go bad by expanding a single event into the whole line $r = 2M$, which would have the effect that if any particle reached the horizon after that event, then it would have to cross $r = 2M$ ‘after’ $t = +\infty$. This singularity would then be very like the one in Fig. 11.9 for spherical coordinates at the pole: a whole line in the bad coordinates representing a point in the real space. Notice that the coordinate diagram in Fig. 11.10 makes no attempt to represent the *geometry* properly, only the coordinates. It clearly does a poor job on the geometry because the light cones close up. Since we have already decided that they *don’t* really close up (particles reach the horizon at finite *proper* time and encounter a perfectly well-behaved geometry there), the remedy is to find coordinates which do not close up the light cones.

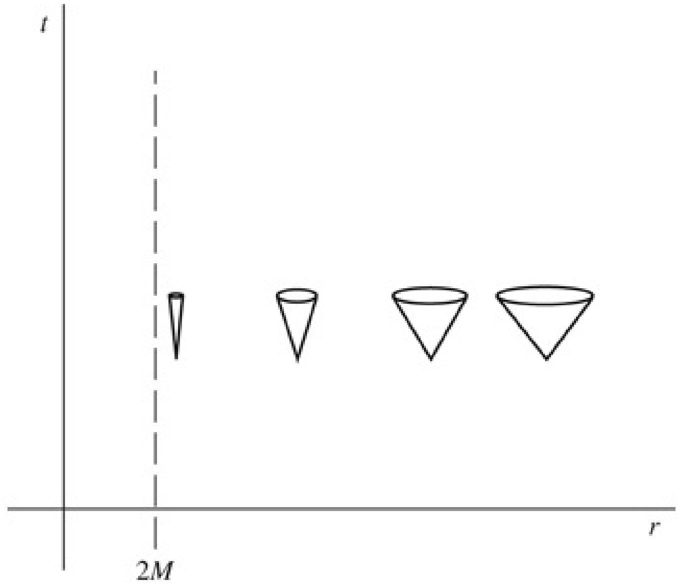


Figure 11.10 Light cones drawn in Schwarzschild coordinates close up near the surface $r = 2M$.

Kruskal–Szekeres coordinates

The search for these coordinates was a long and difficult one, and ended in 1960. The good coordinates are known as Kruskal–Szekeres coordinates, are called u and v, and are defined by

$$\left.\begin{aligned} u &= (r/2M - 1)^{1/2} e^{r/4M} \cosh(t/4M), \\ v &= (r/2M - 1)^{1/2} e^{r/4M} \sinh(t/4M), \end{aligned}\right\} \tag{11.65}$$

for $r > 2M$ and

$$\left.\begin{aligned} u &= (1 - r/2M)^{1/2} e^{r/4M} \sinh(t/4M), \\ v &= (1 - r/2M)^{1/2} e^{r/4M} \cosh(t/4M), \end{aligned}\right\} \tag{11.66}$$

for $r < 2M$. (This transformation is singular at $r = 2M$, but that is necessary in order to eliminate the coordinate singularity there.) The metric in these coordinates is found to be

$$\mathrm{d}s^2 = -\frac{32M^3}{r} e^{-r/2M} (\mathrm{d}v^2 - \mathrm{d}u^2) + r^2 \mathrm{d}\Omega^2, \tag{11.67}$$

where, now, r is not to be regarded as a coordinate but as a function of u and v, given implicitly by the inverse of Eqs. (11.65) and (11.66):

$$\left(\frac{r}{2M} - 1\right) e^{r/2M} = u^2 - v^2. \tag{11.68}$$

Notice several things about Eq. (11.67). There is nothing singular about any metric term at $r = 2M$. There *is*, however, a singularity at $r = 0$, where we expect it. A radial null line ($\mathrm{d}\theta = \mathrm{d}\phi = \mathrm{d}s = 0$) is a line

$$\mathrm{d}v = \pm \mathrm{d}u. \tag{11.69}$$

This last result is very important. It means that in a (u, v) diagram, the light cones are all as open as in SR. This result makes these coordinates particularly useful for visualizing the geometry in a coordinate diagram. The (u, v) diagram is, then, given in Fig. 11.11. Compare this with the result of Exer. 21, § 5.8.

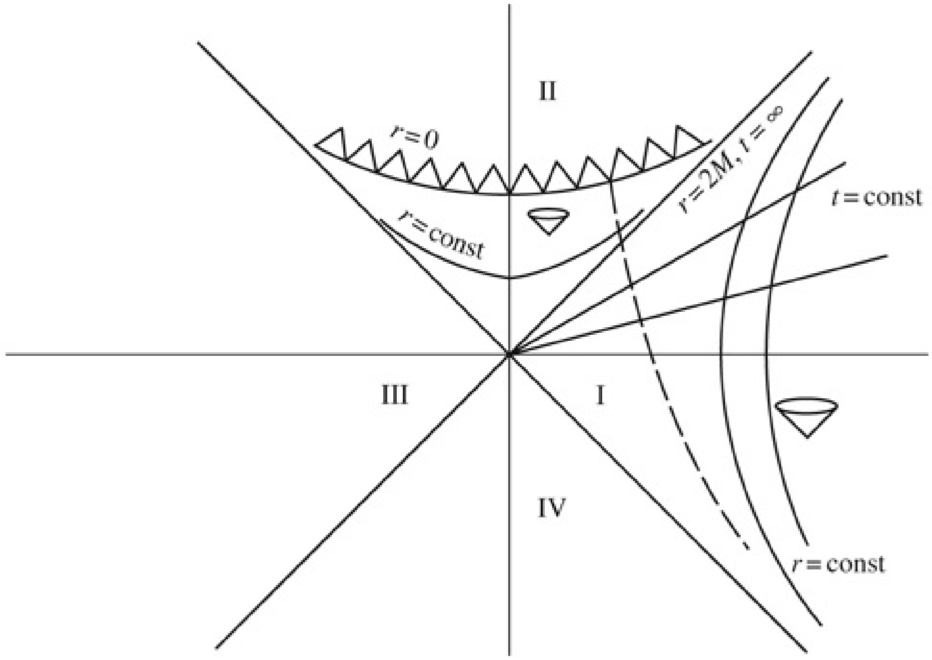


Figure 11.11 Kruskal–Szekeres coordinates keep the light cones at 45° everywhere. The singularity at $r = 0$ (toothed line) bounds the future of all events inside (above) the line $r = 2M$, $t = +\infty$. Events outside this horizon have part of their future free of singularities.

Much needs to be said about this. First, two light cones are drawn for illustration. *Any* 45° line is a radial null line. Second, only u and υ are plotted: θ and ϕ are suppressed; therefore each point is really a two sphere of events. Third, lines of constant r are hyperbolae, as is clear from Eq. (11.68). For $r > 2M$, these hyperbolae run roughly vertically, being asymptotic to the 45° line from the origin $u = \upsilon = 0$. For $r < 2M$, the hyperbolae run roughly horizontally, with the same asymptotes. This means that for $r < 2M$, a timelike line (confined within the light cone) cannot remain at constant r. This is the result we had before. The hyperbola $r = 0$ is the end of the spacetime, since a true singularity is there. Note that although $r = 0$ is a 'point' in ordinary space, it is a whole hyperbola here. However, not too much can be made of this, since it is a singularity of the geometry: we should not glibly speak of it as a part of spacetime with a well-defined dimensionality.

Our fourth remark is that lines of constant t, being orthogonal to lines of constant r, are straight lines in this diagram, radiating outwards from the origin $u = \upsilon = 0$. (They are orthogonal to the hyperbolae $r =$ const. in the *spacetime* sense of orthogonality; recall our diagrams in § 1.7 of invariant hyperbolae in SR, which had the same property of being orthogonal to lines radiating out from the origin.) In the limit as $t \to \infty$, these lines approach the 45° line from the origin. Since all the lines t = const. pass through the origin, the origin would be expanded into a whole line in a (t, r) coordinate diagram like Fig. 11.10, which is what we guessed after discussing that diagram. A world line crossing this $t = \infty$ line in Fig. 11.11 enters the region in which r is a time coordinate, and so cannot get out again. The true horizon, then, is this line $r = 2M$, $t = +\infty$.

Fifth, since for a distant observer t really does measure proper time, and an object that falls to the horizon crosses all the lines t = const. up to $t = \infty$, a distant observer would conclude that it takes an infinite time for the infalling object to reach the horizon. We have already drawn this conclusion before, but here we see it displayed clearly in the diagram. There is nothing 'wrong' in this statement: the distant observer does wait an infinite time to get the information that the object has crossed the horizon. But the object reaches the horizon in a finite time on its own clock. If the infalling object sends out radio pulses each time its clock ticks, then it will emit only a finite number before reaching the horizon, so the distant observer can receive only a finite number of pulses. Since these are stretched out over a very large amount of the distant observer's time, the observer concludes that time on the infalling clock is slowing down and eventually stopping. If the infalling 'clock' is a photon, the observer will conclude that the photon experiences an infinite gravitational redshift. This will also happen if the infalling 'object' is a gravitational wave of short wavelength compared to the horizon size.

Sixth, this horizon is itself a null line. This *must* be the case, since the horizon is the boundary between null rays that cannot get out and those that can. It is therefore the path of the 'marginal' null ray. Seventh, the 45° lines from the origin divide spacetime up into four regions, labeled I, II, III, IV. Region I is clearly the 'exterior', $r > 2M$, and region II is the interior of the horizon. But what about III and IV? To discuss them is beyond the scope of this publication (see Misner *et al.* 1973, Box 33.2G and Ch. 34; and Hawking and Ellis 1973), but one remark must be made. Consider the dashed line in Fig. (11.11), which could be the path of an infalling particle. If this black hole were formed by the collapse of a star, then we know that outside the star the geometry is the Schwarzschild geometry, but inside it may be quite different. The dashed line may be taken to be the path of the *surface* of the collapsing star, in which case the region of the diagram to the right of it is *outside* the star and so correctly describes the spacetime geometry, but everything to the left would be inside the star (smaller r) and hence has possibly no relation to the true geometry of the spacetime. This includes all of regions III and IV, so they are to be ignored by the astrophysicist (though they can be interesting to the mathematician!). Note that parts of I and II are also to be ignored, but there is still a singularity and horizon outside the star.

The eighth and last remark we will make is that the coordinates u and υ are *not* particularly good for describing the geometry far from the star, where g_{uu} and $g_{\upsilon\upsilon}$ fall off exponentially in r. The coordinates t and r are best there; indeed, they were constructed in order to be well behaved there. But if we is interested in the horizon, then we use u and v.

11.3 General black holes

Formation of black holes in general

The phenomenon of the formation of a horizon has to do with the collapse of matter to such small dimensions that the gravitational field traps everything within a certain region, which is called the interior of the horizon. We have explored the structure of the black hole in one particular case – the static, spherically symmetric situation – but the formation of a horizon is a much more general phenomenon. When we discuss astrophysical black holes in § 11.4 we will address the question of how Nature might arrange to get so much mass into such a small region. But it should be clear that, in the real world, black holes are not formed from perfectly spherical collapse. Black holes form in complicated dynamical circumstances, and after they form they continue to participate in dynamics: as members of binary systems, as centers of gas accretion, as sources of gravitational radiation. In this section we learn how to define black holes and what we know about them in general.

The central, even astonishing property of the Schwarzschild horizon is that anything that crosses it can not get back outside it. The definition of a general horizon (called an event horizon) focuses on this property. *An event horizon is the boundary in spacetime between events that can communicate with distant observers and events that cannot.* This definition assumes that distant observers exist, that the spacetime is asymptotically flat, as defined in § 8.4. And it permits the communication to take an arbitrarily long time: an event is considered to be outside the horizon provided it can emit a photon in even just one special direction that eventually makes it out to a distant observer. The most important part of the definition to think about is that the horizon is a boundary in *spacetime,* not just in the space defined by one moment of time. It is a three-dimensional surface that separates the events of spacetime into two regions: trapped events inside the horizon and untrapped events outside.

Since no form of communication can go faster than light, the test of whether events can communicate with distant observers is whether they can send light rays, that is whether there are null rays that can get arbitrarily far away. As the boundary between null rays that can escape and null rays that are trapped, the horizon itself is composed of null world lines. These are the marginally trapped null rays, the ones that neither move away to infinity nor fall inwards. By definition these marginal null rays stay on the horizon forever, because if a ray were to leave it toward the exterior or interior, then it would not mark the horizon. It is not hard to see that this definition fits the Schwarzschild horizon, which is static and unchanging, but when we consider dynamical situations there are some surprises.

The formation of a horizon from a situation where there is initially no black hole illustrates well the dynamical nature of the horizon. Consider the collapse of a spherical star to form a black hole. In the end there is a static Schwarzschild horizon, but before that there is an intermediate period of time in which the horizon is growing from zero radius to its full size. This is easy to see by considering Fig. 11.12, which illustrates (*very* schematically) the collapsing situation. (The time coordinate is a kind of Schwarzschild time, but it isn't to be taken too literally.) As matter falls in, the trajectories of photons that start from the center of the collapsing star (wavy lines) are more and more affected. Photon (*a*) gets out with little trouble, photon (*b*) has some delay, and photon (*c*) is the 'marginal' one, which just gets trapped and remains on the Schwarzschild horizon. Anything later than (*c*) is permanently trapped, anything earlier gets out. So photon (*c*) does in fact represent the horizon at all times, by definition, since it is the boundary between trapped and untrapped. Thus, we see the horizon grow from zero radius to $2M$ by watching photon (*c*)'s progress outwards. For this spherically symmetric situation, if we knew the details of the collapse, we could easily determine the position of the horizon. But if there were no symmetry – particularly if the collapse produced a large amount of gravitational radiation – then the calculation would be far more difficult, although the principle would be the same.

As a more subtle example of a dynamical horizon, consider what happens to the Schwarzschild horizon of the 'final' black hole in Fig. 11.12 if at a much later time some more gas falls in (not illustrated in the figure). For simplicity, we again assume the infalling gas is perfectly spherical. Let the mass of the hole before the gas falls in be called M_0. The surface $r = 2M_0$ is static and appears to be the event horizon: photons inside it fall towards the singularity at the center, and photons infinitesimally outside it gradually move further and further away from it. But then the new gas falls across this surface and increases the mass of the hole to M_1. Clearly, the new final state will be a Schwarzschild solution of mass M_1, where the larger surface $r = 2M_1$ looks like an event horizon. It consists of null rays neither falling in nor diverging outwards. Now, what is the history of these rays? What happens if we trace backwards in time along a ray that just stays at $r = 2M_1$? We would find that, before the new gas arrived, it was one of those null world lines just *outside* the surface $r = 2M_0$, one of the rays that were very gradually diverging from it. The extra mass has added more gravitational attraction (more curvature) that stopped the ray from moving away and now holds it exactly at $r = 2M_1$. Moreover, the null rays that formed the static surface $r = 2M_0$ before the gas fell in are now inside $r = 2M_1$ and are therefore falling toward the singularity, again pulled in by the gravitational attraction of the extra mass. The boundary in *spacetime* between what is trapped and what is not consists therefore of the null rays that in the end sit at $r = 2M_1$, including their continuation backwards in time. The null rays on $r = 2M_0$ were not actually part of the true horizon of spacetime even at the earlier time: they are just trapped null rays that took a long time to find out that they were trapped! Before the gas fell in, what looked like a static event horizon ($r = 2M_0$) was not an event horizon at all, even though it was temporarily a static collection of null rays. Instead, the true boundary between trapped and untrapped was even at that early time gradually expanding outwards, traced by the null world lines that eventually became the surface $r = 2M_1$.

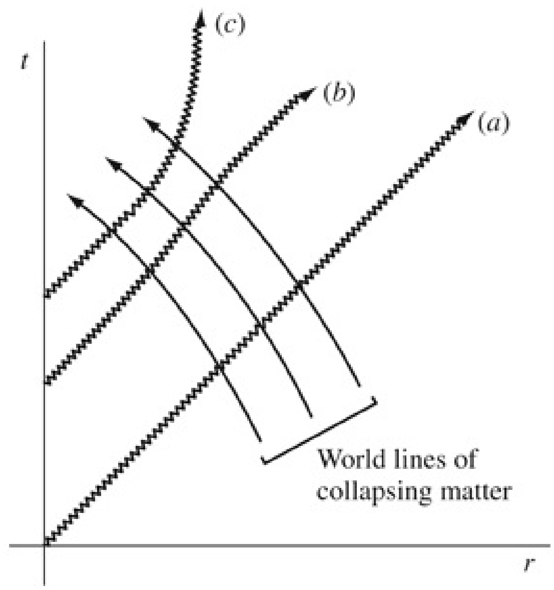


Figure 11.12 Schematic spacetime diagram of spherical collapse. Light ray (*a*) hardly feels anything, (*b*) is delayed, and (*c*) is marginally trapped. The horizon is defined as the ray (*c*), so it grows continuously from zero radius as the collapse proceeds.

This example illustrates the fact that the horizon is not a location in space but a boundary in spacetime. It is not possible to determine the location of the horizon by looking at a system at one particular time; instead we must look at its entire evolution in time, find out which null rays eventually really do escape and which ones are trapped, and then trace out the boundary between them. The horizon is a property of the spacetime as a whole, not of space at any one moment of time. In our example, if many years later some further gas falls into the hole, then we would find that even those null rays at $r = 2M_1$ would actually be trapped and would not have been part of the true event horizon after all. The only way to find the true horizon is to know the *entire* future evolution of the spacetime and then trace out the boundary between trapped and untrapped regions.

This definition of the event horizon is mathematically consistent and logical, but it is difficult to work with in practical situations. If we know only a limited amount about a spacetime, it can in principle be impossible to locate the horizon. For example, computer simulations of solutions of Einstein's equations containing black holes cannot run forever, so they don't have all the information needed even to locate the horizons of their black holes. So physicists also work with another kind of surface that can be defined at any one moment of time. It uses the other property of the Schwarzschild horizon, that it is a static collection of null rays. Physicists define a *locally trapped surface* to be a two-dimensional surface at any particular time whose outwardly directed null rays are neither expanding nor contracting at that particular moment. Locally trapped surfaces are always inside true horizons in dynamical situations, just as $r = 2M_0$ turned out to be inside the true horizon, but very often they are so close to the true horizon that they offer an excellent approximation to it.

General properties of black holes

While the detailed structure of an event horizon is not easy to compute, some important general properties of horizons are understood, and they underpin the confidence with which astronomers now employ black holes in models of complex astrophysical phenomena. Here are some important theorems and conjectures.

) It is believed that any horizon will eventually become stationary, provided that it is not constantly disturbed by outside effects like accretion. So an isolated black hole should become stationary. (The calculations in item (3) below support this idea.) The *stationary* horizons, by contrast with nonstationary ones, are completely known. The principal result is that a stationary vacuum black hole is characterized by just two numbers: its total mass M and total angular momentum J. These parameters are defined not by any integrals over the ‘interior’ of the horizon, but by the gravitational field far from the hole. We have defined the mass M of any metric in this fashion in § 8.4, and in Exer. 19, § 8.6, we have seen how J can be similarly defined. The unique stationary vacuum black hole is the Kerr solution, Eq. (11.71), which we study in detail below. If the angular momentum is zero, the Kerr solution becomes the Schwarzschild metric. This uniqueness theorem results from work done by Hawking (1972), Carter (1973), and Robinson (1975). See Chruściel (1996) and Heusler (1998) for reviews.

) If the black hole is not in vacuum, its structure may be more general. It may carry an electric charge Q and, in principle, a magnetic monopole moment F, although magnetic monopoles are not found in Nature. Both of these charges can be measured by Gauss’ law integrals over surfaces surrounding the hole and far from it. It is also felt that collapse is unlikely to lead to a significant residual charge Q, so astrophysicists normally take only M and J to be nonzero. But other kinds of fields, such as those encountered in particle physics theories, can add other complications: self-gravitating interacting scalar fields, Yang-Mills fields, and so on. See Heusler (1998) for a review. Again, these complications are not usually thought to be relevant in astrophysics. What may be relevant, however, is the distortion of the horizon produced by the tidal effect of matter surrounding the black hole. If a massive stationary disk of gas surrounds the hole, then the metric will not be exactly Kerr (e.g. Will 1974, 1975; Ansorg 2005).

) If gravitational collapse is *nearly* spherical, then all nonspherical parts of the mass distribution – quadrupole moment, octupole moment – except possibly for angular momentum, are radiated away in gravitational waves, and a stationary black hole of the Kerr type is left behind. If there is no angular momentum, a Schwarzschild hole is left behind (Price 1972a, b). The Kerr black hole appears to be stable against all perturbations, but the proof is not quite complete (Whiting 1989, Beyer 2001).

) An important general result concerning nonstationary horizons is the area theorem of Hawking (Hawking and Ellis 1973): in any dynamical process involving black holes, the total area of all the horizons cannot decrease in time. We saw this in a qualitative way in our earlier discussion of how to define a horizon when matter is falling into a black hole: the area is actually increasing during the period before the infalling gas reaches the hole. We shall see below how to quantify this theorem by calculating the area of the Kerr horizon. The area theorem implies that, while two black holes can collide and coalesce, a single black hole can never bifurcate spontaneously into two smaller ones. (A restricted proof of this is in Exer. 26, § 11.7 using the Kerr area formula below; a full proof is outlined in Misner *et al.* 1973, Exer. 34.4, and requires techniques beyond the scope of this book.) The theorem assumes that the local energy density of matter in spacetime (ρ) is positive. The analogy between an ever-increasing area and the ever-increasing entropy of thermodynamics has led to the development of black-hole thermodynamics, and the understanding that black holes fit into thermal physics in a very natural way. The entropy associated with the area of the horizon is given in Eq. (11.114). The key to this association is the demonstration by Hawking that quantum mechanics can lead to spontaneous radiation from a black hole. The radiation has a thermal spectrum, which leads to a temperature and hence an entropy for the horizon. This violation of the area theorem by quantum effects happens because in quantum mechanics, energies are not always required to be positive. We study this Hawking radiation in § 11.5 below.

) Inside the Schwarzschild and Kerr horizons there are curvature singularities where the curvature, and hence the tidal gravitational force, becomes infinite. General theorems, mostly due to Hawking and Penrose, imply that any horizon will contain a singularity within it of some kind (Hawking and Ellis 1973). But it is not known whether such singularities will always have infinite curvature; all that is known is that infalling geodesics are incomplete and cannot be continued for an infinite amount of proper time or affine distance. The existence of these singularities is generally regarded as a serious shortcoming of general relativity, that its predictions have limited validity in time inside horizons. Many physicists expect this shortcoming to be remedied by a quantum theory of gravity. The uncertainty principle of quantum mechanics, so the expectation goes, will make the singularity a little 'fuzzy', the tidal forces will not quite reach infinite strength, and the waveform will continue further into the future. In the absence of an acceptable theory of quantum gravity, this remains only a hope, but there are some computations in restricted quantum models that are suggestive that this might in fact happen (Bojowald 2005, Ashtekar and Bojowald 2006).

) The generic existence of singularities inside horizons, hidden from the view of distant observers, prompts the question of whether there can be so-called *naked singularities*, that is singularities outside horizons. These would be far more problematic for general relativity, for it would mean that situations could arise in which general relativity could make no predictions beyond a certain time even for normal regions of spacetime. Having singularities in unobservable regions inside horizons is bad enough, but if singularities arose outside horizons, general relativity would be even more flawed. In response to this concern, Penrose (1979) formulated the *cosmic censorship conjecture*, according to which no naked singularities can arise out of nonsingular initial conditions in asymptotically flat spacetimes. Penrose had no proof of this, and offered it as a challenge to relativity theorists. Even today there is considerable debate over whether, and in exactly what formulation, the conjecture is true (Berger 2002, Cruściel 1991, Rendall 2005). One naked singularity seems inescapable in general relativity: the Big Bang of standard cosmology is naked to our view. If the universe re-collapses, there will similarly be a Big Crunch in the future of all our world lines. These issues are discussed in the next chapter.

Item (1) above is truly remarkable: a massive black hole, possibly formed from 10^{60} individual atoms and molecules – whose history as a gas may have included complex gas motions, shock waves, magnetic fields, nucleosynthesis, and all kinds of other complications – is described fully and exactly by just two numbers, its mass and spin. The horizon and the entire spacetime geometry outside it depend on just these two numbers. All the complication of the formation process is effaced, forgotten, reduced to two simple numbers. No other macroscopic body is so simple to describe. We might characterize a star by its mass, luminosity, and color, but these are just a start, just categories that contain an infinite potential variety within them. Stars can have magnetic fields, spots, winds, differential rotation, and many other largescale features, to say nothing of the different motions of individual atoms and ions. While this variety may not be relevant in most circumstances, it is there. For a black hole, it is simply not there. There is nothing but mass and spin, no individual structure or variety revealed by microscopic examination of the horizon (see the reaction of Chandrasekhar to this fact, quoted in § 11.4 below). In fact, the horizon is not even a real surface, it is just a boundary in empty space between trapped and untrapped regions. The fact that no information can escape from inside the hole means that no information about what fell in is visible from the outside. The only quantities that remain are those that are conserved by the fundamental laws of physics: mass and angular momentum.

Kerr black hole

The Kerr black hole is axially symmetric but not spherically symmetric (that is rotationally symmetric about one axis only, which is the angular-momentum axis), and is characterized by two parameters, M and J. Since J has dimension m^2, we conventionally define

$$a := J/M, \tag{11.70}$$

which then has the same dimensions as M. The line element is

$$\begin{aligned} \mathrm{d}s^2 = &- \frac{\Delta - a^2 \sin^2\theta}{\rho^2}\mathrm{d}t^2 - 2a\frac{2Mr\sin^2\theta}{\rho^2}\mathrm{d}t\,\mathrm{d}\phi \\ &+ \frac{(r^2+a^2)^2 - a^2\Delta\sin^2\theta}{\rho^2}\sin^2\theta\,\mathrm{d}\phi^2 + \frac{\rho^2}{\Delta}\mathrm{d}r^2 + \rho^2\,\mathrm{d}\theta^2, \end{aligned} \tag{11.71}$$

where

$$\begin{aligned} \Delta &:= r^2 - 2Mr + a^2, \\ \rho^2 &:= r^2 + a^2\cos^2\theta. \end{aligned} \tag{11.72}$$

The coordinates are called Boyer–Lindquist coordinates; ϕ is the angle around the axis of symmetry, t is the time coordinate in which everything is stationary, and r and θ are similar to the spherically symmetric r and θ but are not so readily associated to any geometrical definition. In particular, since there are no metric two-spheres, the coordinate r cannot be defined as an 'area' coordinate as we did before. The following points are important:

) Surfaces t = const., r = const. do not have the metric of the two-sphere, Eq. (10.2).
) The metric for $a = 0$ is identically the Schwarzschild metric.
) There is an off-diagonal term in the metric, in contrast to Schwarzschild:

$$g_{t\phi} = -a\frac{2Mr\sin^2\theta}{\rho^2}, \tag{11.73}$$

which is $\frac{1}{2}$ the coefficient of $dt\ d\phi$ in Eq. (11.71) because the line element contains two terms,

$$g_{t\phi}\ dt\ d\phi + g_{\phi t}\ d\phi\ dt = 2g_{t\phi}\ d\phi\ dt,$$

by the symmetry of the metric. *Any* axially symmetric, stationary metric has preferred coordinates t and ϕ, namely those which have the property $g_{\alpha\beta,t} = 0 = g_{\alpha\beta,\phi}$. But the coordinates r and θ are more-or-less arbitrary, except that they may be chosen to be (i) orthogonal to t and ϕ ($g_{rt} = g_{r\phi} = g_{\theta t} = g_{\theta\phi} = 0$) and (ii) orthogonal to each other ($g_{\theta r} = 0$). In general, we cannot choose t and ϕ orthogonal to each other ($g_{t\phi} \neq 0$). Thus Eq. (11.71) has the minimum number of nonzero $g_{\alpha\beta}$ (see Carter 1969).

Dragging of inertial frames

The presence of $g_{t\phi} \neq 0$ in the metric introduces qualitatively new effects on particle trajectories. Because $g_{\alpha\beta}$ is independent of ϕ, a particle's trajectory still conserves p_ϕ. But now we have

$$p^\phi = g^{\phi\alpha}p_\alpha = g^{\phi\phi}p_\phi + g^{\phi t}p_t, \tag{11.74}$$

and similarly for the time components:

$$p^t = g^{t\alpha}p_\alpha = g^{tt}p_t + g^{t\phi}p_\phi. \tag{11.75}$$

Consider a zero angular-momentum particle, $p_\phi = 0$. Then, using the definitions (for nonzero rest mass)

$$p^t = m\,\mathrm{d}t/\mathrm{d}\tau, \quad p^\phi = m\,\mathrm{d}\phi/\mathrm{d}\tau, \tag{11.76}$$

we find that the particle's trajectory has

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = \frac{p^\phi}{p^t} = \frac{g^{\phi t}}{g^{tt}} := \omega(r,\theta). \tag{11.77}$$

This equation defines what we mean by ω, the angular velocity of a zero angular-momentum particle. We shall find ω explicitly for the Kerr metric when we obtain the contravariant components $g^{\phi t}$ and g^{tt} below. But it is clear that this effect will be present in any metric for which $g_{t\phi} \neq 0$, which in turn happens whenever the source is rotating (e.g. a rotating star as in Exer. 19, § 8.6). So we have the remarkable result that a particle dropped 'straight in' ($p_\phi = 0$) from infinity is 'dragged' just by the influence of gravity so that it acquires an angular velocity in the same sense as that of the source of the metric (we'll see below that, for the Kerr metric, ω has the same sign as a). This effect weakens with distance (roughly as $1/r^3$; see Eq. (11.90) below for the Kerr metric), and it makes the angular momentum of the source measurable in principle, although in most situations the effect is small, as we have seen in Exer. 19, § 8.6. This effect is often called the *dragging of inertial frames.*

This effect has a close analogy with magnetism. Newtonian gravity is, of course, very similar to electrostatics, with the sign change that ensures that 'charges' of the same sign attract one another. In electromagnetism, a spinning charge creates additional effects which we call magnetism. Here we have the gravitational analog, the gravitational effects due to a spinning mass. For that reason these effects are called *gravitomagnetism*. The analogy between gravitomagnetism and standard magnetism is perhaps easier to see in the *Lense–Thirring effect:* a gyroscope placed in orbit around a rotating star will precess by a small amount that is proportional to the angular momentum of the star, just as a spinning electron precesses if it orbits through a magnetic field.

Small as it is, the Lense–Thirring effect created by the spin of the Earth has been measured. Detailed studies of the orbits of two geodesy satellites have so far been able to verify that the effect is as predicted by general relativity, with an accuracy at the 20% level (Ciufolini *et al.*, 2006). An order of magnitude improvement could be made with an additional satellite in a specially chosen orbit. In addition, a satellite experiment called Gravity Probe B (GP-B), which has measured the precession of on-board gyroscopes, is currently (2008) analyzing its data and is expected to report soon, hopefully with errors below the 10% level. GP-B is one of the most sensitive and high-precision experiments ever performed in orbit. Gravitomagnetic effects are regularly taken into account in modeling the details of the emission of X-rays near black holes (e.g. Brenneman and Reynolds 2006), and they may also soon be seen in the double pulsar system PSR J0737-3039 referred to earlier in this chapter.

Ergoregion

Consider photons emitted in the equatorial plane ($\theta = \pi/2$) at some given r. In particular, consider those initially going in the $\pm\phi$-direction, that is tangent to a circle of constant r. Then they generally have only dt and dϕ nonzero on the path at first and since d$s^2 = 0$, we
have

$$\begin{aligned} 0 &= g_{tt}\,\mathrm{d}t^2 + 2g_{t\phi}\,\mathrm{d}t\,\mathrm{d}\phi + g_{\phi\phi}\,\mathrm{d}\phi^2 \\ \Rightarrow \frac{\mathrm{d}\phi}{\mathrm{d}t} &= -\frac{g_{\tau\phi}}{g_{\phi\phi}} \pm \left[\left(\frac{g_{t\phi}}{g_{\phi\phi}}\right)^2 - \frac{g_{tt}}{g_{\phi\phi}}\right]^{1/2}. \end{aligned} \tag{11.78}$$

Now, a remarkable thing happens if $g_{tt} = 0$: the two solutions are

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = 0 \quad \text{and} \quad \frac{\mathrm{d}\phi}{\mathrm{d}t} = -\frac{2g_{t\phi}}{g_{\phi\phi}}. \tag{11.79}$$

We will see below that for the Kerr metric the second solution gives dϕ/dt the *same* sign as the parameter a, and so represents the photon sent off in the same direction as the hole's rotating. The other solution means that the other photon – the one sent 'backwards'– initially doesn't move at all. The dragging of orbits has become so strong that this photon *cannot* move in the direction opposite the rotation. Clearly, any particle, which must move slower than a photon, will therefore have to rotate with the hole, even if it has an angular momentum arbitrarily large in the opposite sense to the hole's!

We shall see that the surface where $g_{tt} = 0$ lies outside the horizon; it is called the ergosphere. It is sometimes also called the 'static limit', since inside it no particle can remain at fixed r, θ, ϕ. From Eq. (11.71) we conclude that it occurs at

$$r_0 := r_{\text{ergosphere}} = M + \surd(M^2 - a^2 \cos^2\theta). \tag{11.80}$$

Inside this radius, since $g_{tt} > 0$, all particles and photons must rotate with the hole.

Again, this effect can occur in other situations. Models for certain rotating stars are known where there are toroidal regions of space in which $g_{tt} > 0$ (Butterworth and Ipser 1976). These will have these super-strong frame-dragging effects. They are called ergoregions, and their boundaries are ergotoroids. They can exist in solutions which have no horizon at all. But the stars are extremely compact (relativistic) and very rapidly rotating. It seems unlikely that real neutron stars would have ergoregions.

The Kerr horizon

In the Schwarzschild solution, the horizon was the place where $g_{tt} = 0$ and $g_{rr} = \infty$. In the Kerr solution, the ergosphere occurs at $g_{tt} = 0$ and the horizon is at $g_{rr} = \infty$, *i.e.* where $\Delta = 0$:

$$r_+ := r_{\text{horizon}} = M + \surd(M^2 - a^2). \tag{11.81}$$

It is clear that the Kerr ergosphere lies outside the horizon except at the poles, where it is tangent to it. The full proof that this is the horizon is beyond our scope here: we need to verify that no null lines can escape from inside r_+. We shall simply take it as given (see the next section below for a partial justification). Since the area of the horizon is important (Hawking's area theorem), we shall calculate it.

The horizon is a surface of constant r and t, by Eq. (11.81) and the fact that the metric is stationary. Any surface of constant r and t has an intrinsic metric whose line element comes from Eq. (11.71) with dt = dr = 0:

$$\mathrm{d}l^2 = \frac{(r^2 + a^2)^2 - a^2\Delta}{\rho^2} \sin^2\theta \, \mathrm{d}\phi^2 + \rho^2 \, \mathrm{d}\theta^2. \tag{11.82}$$

The proper area of this surface is given by integrating the square root of the determinant of this metric over all θ and ϕ:

$$A(r) = \int_0^{2\pi} \mathrm{d}\phi \int_0^{\pi} \mathrm{d}\theta \sqrt{[(r^2 + a^2)^2 - a^2 \Delta]} \sin\theta. \tag{11.83}$$

Since nothing in the square root depends on θ or ϕ, and since the area of a unit two-sphere is

$$4\pi = \int_0^{2\pi} \mathrm{d}\phi \int_0^{\pi} \mathrm{d}\theta \ \sin\theta,$$

we immediately conclude that

$$A(r) = 4\pi \sqrt{[(r^2 + a^2)^2 - a^2 \Delta]}. \tag{11.84}$$

Since the horizon is defined by $\Delta = 0$, we get

$$A(\text{horizon}) = 4\pi (r_+^2 + a^2). \tag{11.85}$$

Equatorial photon motion in the Kerr metric

A detailed study of the motion of photons in the equatorial plane gives insight into the ways in which 'rotating' metrics differ from nonrotating ones. First, we must obtain the inverse of the metric, Eq. (11.71), which we write in the general stationary, axially symmetric form:

$$\mathrm{d}s^2 = g_{tt}\,\mathrm{d}t^2 + 2g_{t\phi}\,\mathrm{d}t\,\mathrm{d}\phi + g_{\phi\phi}\,\mathrm{d}\phi^2 + g_{rr}\,\mathrm{d}r^2 + g_{\theta\theta}\,\mathrm{d}\theta^2.$$

The only off-diagonal element involves t and ϕ; therefore

$$g^{rr} = \frac{1}{g_{rr}} = \Delta\rho^{-2}, \quad g^{\theta\theta} = \frac{1}{g_{\theta\theta}} = \rho^{-2}. \tag{11.86}$$

We need to invert the matrix

$$\begin{pmatrix} g_{tt} & g_{t\phi} \\ g_{t\phi} & g_{\phi\phi} \end{pmatrix}.$$

Calling its determinant D, the inverse is

$$\frac{1}{D}\begin{pmatrix} g_{\phi\phi} & -g_{t\phi} \\ -g_{t\phi} & g_{tt} \end{pmatrix}, \quad D = g_{tt}g_{\phi\phi} - (g_{t\phi})^2. \tag{11.87}$$

Notice one important deduction from this. The angular velocity of the dragging of inertial frames is Eq. (11.77):

$$\omega = \frac{g^{\phi t}}{g^{tt}} = \frac{-g_{t\phi}/D}{g_{\phi\phi}/D} = -\frac{g_{t\phi}}{g_{\phi\phi}}. \tag{11.88}$$

This makes Eqs. (11.78) and (11.79) more meaningful. For the metric, Eq. (11.71), some algebra gives

$$\begin{aligned} &D = -\Delta \sin^2\theta, \quad g^{tt} = -\frac{(r^2+a^2)^2 - a^2\Delta\sin^2\theta}{\rho^2\Delta}, \\ &g^{t\phi} = -a\frac{2Mr}{\rho^2\Delta}, \quad g^{\phi\phi} = \frac{\Delta - a^2\sin^2\theta}{\rho^2\Delta\sin^2\theta}. \end{aligned} \tag{11.89}$$

Then the frame dragging is

$$\omega = \frac{2Mra}{(r^2+a^2)^2 - a^2\Delta\sin^2\theta}. \tag{11.90}$$

The denominator is positive everywhere (by Eq. (11.72)), so this has the same sign as a, and it falls off for large r as r^{-3}, as we noted earlier.

A photon whose trajectory is in the equatorial plane has $d\theta = 0$; but, unlike the Schwarzschild case, this is only a special kind of trajectory: photons not in the equatorial plane may have qualitatively different orbits. Nevertheless, a photon for which $p^\theta = 0$ initially in the equatorial plane always has $p^\theta = 0$, since the metric is reflection symmetric through the plane $\theta = \pi/2$. By stationarity and axial symmetry the quantities $E = -p_t$ and $L = p_\phi$ are constants of the motion. Then the equation $\vec{p}\cdot\vec{p} = 0$ determines the motion. Denoting p^r by $dr/d\lambda$ as before, we get, after some algebra,

$$\begin{aligned} \left(\frac{dr}{d\lambda}\right)^2 &= g^{rr}[(-g^{tt})E^2 + 2g^{t\phi}EL - g^{\phi\phi}L^2] \\ &= g^{rr}(-g^{tt})\left[E^2 - 2\,\omega EL + \frac{g^{\phi\phi}}{g^{tt}}L^2\right]. \end{aligned} \tag{11.91}$$

Using Eqs. (11.72), (11.86), and (11.90) for $\theta = \pi/2$, we get

$$\left(\frac{\mathrm{d}r}{\mathrm{d}\lambda}\right)^2 = \frac{(r^2+a^2)^2 - a^2\Delta}{r^4}\left[E^2 - \frac{4Mra}{(r^2+a^2)^2 - a^2\Delta}EL - \frac{r^2-2Mr}{(r^2+a^2)^2 - a^2\Delta}L^2\right]. \tag{11.92}$$

This is to be compared with Eq. (11.12), to which it reduces when $a = 0$. Apart from the complexity of the coefficients, Eq. (11.92) differs from Eq. (11.12) in a qualitative way in the presence of a term in EL. So we cannot simply define an effective potential V^2 and write $(\mathrm{d}r/\mathrm{d}\lambda)^2 = E^2 - V^2$. What we can do is nearly as good. We can factor Eq. (11.91):

$$\left(\frac{dr}{d\lambda}\right)^2 = \frac{(r^2+a^2)^2 - a^2\Delta}{r^4}(E - V_+)(E - V_-). \tag{11.93}$$

Then $V_\pm$, by Eqs. (11.91) and (11.92), are

$$V_\pm(r) = [\omega \pm (\omega^2 - g^{\phi\phi}/g^{tt})^{1/2}]L \tag{11.94}$$

$$= \frac{2Mra \pm r^2\Delta^{1/2}}{(r^2+a^2)^2 - a^2\Delta}L. \tag{11.95}$$

This is the analog of the *square root* of Eq. (11.14), to which it reduces when $a = 0$. Now, the square root of Eq. (11.14) becomes imaginary inside the horizon; similarly, Eq. (11.95) is complex when $\Delta < 0$. In each case the meaning is that in such a region there are *no* solutions to $\mathrm{d}r/\mathrm{d}\lambda = 0$, *no* turning points regardless of the energy of the photon. Once a photon crosses the line $\Delta = 0$ it *cannot* turn around and get back outside that line. Clearly, $\Delta = 0$ marks the *horizon* in the equatorial plane. What we haven't shown, but what is also true, is that $\Delta = 0$ marks the horizon for trajectories not in the equatorial plane.

We can discuss the qualitative features of photon trajectories by plotting $V_\pm(r)$. We choose first the case $aL > 0$ (angular momentum in the same sense as the hole), and of course we confine attention to $r \geqslant r_+$ (outside the horizon). Notice that for large r the curves (in Fig. 11.13) are asymptotic to zero, falling off as $1/r$. This is the regime in which the rotation of the hole makes almost no difference. For small r we see features not present without rotation: V_- goes through zero (easily shown to be at r_0, the location of the ergosphere) and meets

V_+ at the horizon, both curves having the value $aL/2Mr_+ := \omega_+ L$, where ω_+ is the value of ω on the horizon. From Eq. (11.93) it is clear that a photon can move only in regions where $E > V_+$ or $E < V_-$. We are used to photons with positive E: they may come in from infinity and either reach a minimum r or plunge in, depending on whether or not they encounter the hump in V_+. There is nothing qualitatively new here. But what of those for which $E \leqslant V_-$? Some of these have $E > 0$. Are they to be allowed?

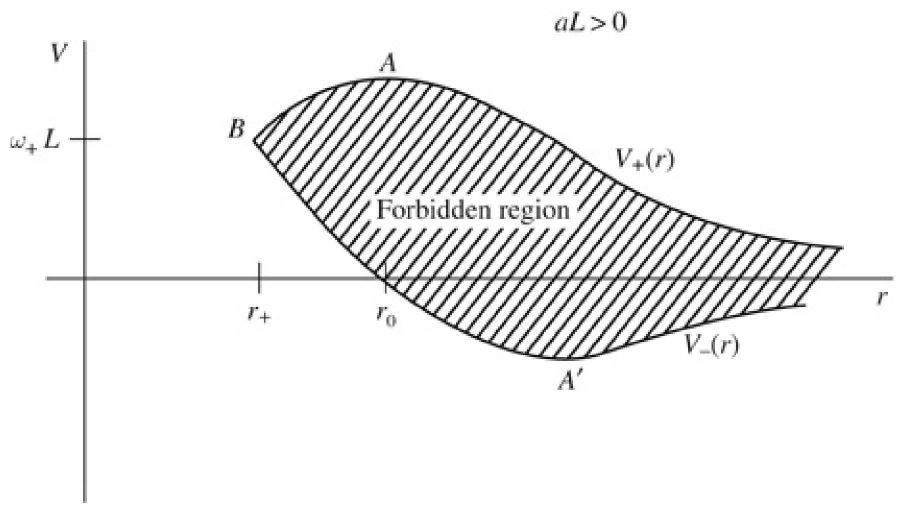


Figure 11.13 Factored potential diagram for equatorial photon orbits of positive angular momentum in the Kerr metric. As $a \to 0$, the upper and lower curves approach the two square roots of Fig. 11.2 outside the horizon.

To discuss negative-energy photons we must digress a moment and talk about moving along a geodesic backwards in time. We have associated our particles' paths with the mathematical notion of a geodesic. Now a geodesic is a curve, and the path of a curve can be traversed in either of two directions; for timelike curves one is forwards and the other backwards in time. The tangents to these two motions are simply opposite in sign, so one will have four-momentum $\vec{p}$ and the other $-\vec{p}$. The energies measured by observer $\vec{U}$ will be $-\vec{U} \cdot \vec{p}$ and $+\vec{U} \cdot \vec{p}$. So one particle will have positive energy and another negative energy. In flat spacetime we conventionally take all particles to travel forwards in time; since all known particles have positive or zero rest mass, this causes them all to have positive energy relative to any Lorentz observer who also moves forwards in time. Conversely, if $\vec{p}$ has positive energy relative to some Lorentz observer, it has positive energy relative to *all* that go forwards in time. In the Kerr metric, however, it will not do simply to demand positive E. This is because E is the energy relative to an observer at infinity; the particle near the horizon is far from infinity, so the direction of 'forward time' isn't so clear. What we must do is set up some observer $\vec{U}$ near the horizon who will have a clock, and demand that $-\vec{p} \cdot \vec{U}$ be positive for particles that pass near him. A convenient observer (but any will do) is one who has zero angular momentum and resides at fixed r, circling the hole at the angular velocity ω. This zero angular-momentum observer (ZAMO) is not on a geodesic, so he must have a rocket engine to remain on this trajectory. (In this respect he is no different from us, who must use our legs to keep us at constant r in Earth's gravitational field.) It is easy to see that he has four-velocity $U^0 = A$, $U^\phi = \omega A$, $U^r = U^\theta = 0$, where A is found from the condition $\vec{U} \cdot \vec{U} = -1 : A^2 = g_{\phi\phi}/(-D)$. This is nonsingular for $r > r_+$. Then he measures the energy of a particle to be

$$\begin{aligned} E_{\mathrm{ZAMO}} &= -\vec{p} \cdot \vec{U} = -(p_0\, U^0 + p_\phi\, U^\phi) \\ &= A(E - \omega L). \end{aligned} \tag{11.96}$$

This is the energy we must demand be positive-definite. Since A is positive, we require

$$E > \omega L. \tag{11.97}$$

From Eq. (11.95) it is clear that any photon with $E > V_+$ also satisfies Eq. (11.97) and so is allowed, while any with $E < V_-$ violates Eq. (11.97) and is moving backwards in time. So in Fig. 11.13 we consider only trajectories for which E lies above V_+; for these there is nothing qualitatively different from Schwarzschild.

The Penrose process

For negative angular-momentum particles, however, new features do appear. If $aL < 0$, it is clear from Eq. (11.95) that the shape of the $V_\pm$ curves is just turned over, so they look like Fig. 11.14. Again, of course, condition Eq. (11.97) means that forward-going photons must lie above $V_+(r)$, but now some of these can have $E < 0$! This happens only for $r < r_0$, *i.e.* inside the ergosphere. Now we see the origin of the name ergoregion: it is from the Greek 'ergo-', meaning energy, a region in which energy has peculiar properties.

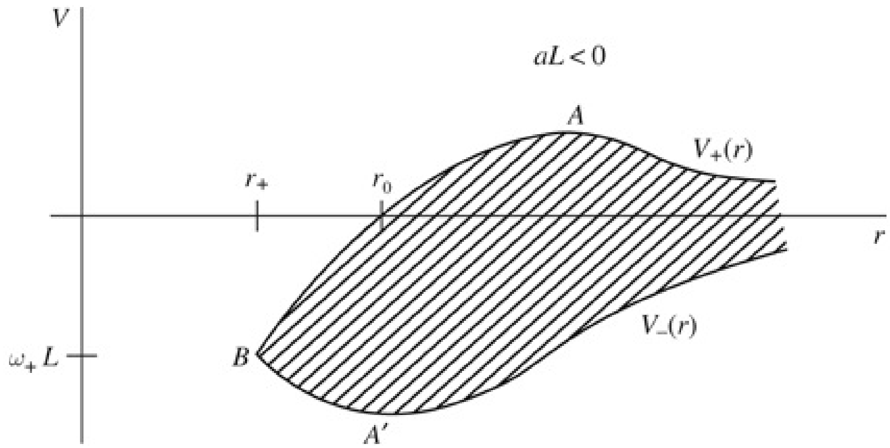


Figure 11.14 As Fig. 11.13 for negative angular momentum photons.

The existence of orbits with negative total energy leads to the following interesting thought experiment, originally suggested by Roger Penrose (1969). Imagine dropping an unstable particle toward a Kerr black hole. When it is inside the ergosphere it decays into two photons, one of which finds itself on a null geodesic with negative energy with respect to infinity. It is trapped in the ergoregion and inevitably falls into the black hole. The other particle must have positive energy, since energy is conserved. If the positive-energy photon can be directed in such a way as to escape from the ergoregion, then the net effect is to leave the hole with less energy than it had to begin with: the infalling particle has 'pumped' the black hole and extracted energy from it! By examining Figs. 11.13 and 11.14, we can convince oneselves that this only works if the negative-energy particle also has a negative L, so that the process involves a decrease in the angular momentum of the hole. The energy has come at the expense of the spin of the black hole.

Blandford and Znajek (1977) suggested that the same process could operate in a practical way if a black hole is surrounded by an accretion disk (see later in this chapter) containing a magnetic field. If the field penetrates inside the ergosphere, then it could facilitate the creation of pairs of electrons and positrons, and some of them could end up on negative-energy trapped orbits in the ergoregion. The escaping particles might form the energetic jets of charged particles that are known to be emitted from near black holes, especially in quasars. In this way, quasar jets might be powered by energy extracted from the rotation of black holes. As of this writing (2008) this is still one of a number of viable competing models for quasar jets.

The Penrose process is not peculiar to the Kerr black hole; it happens whenever there is an ergoregion. If a rotating star has an ergoregion without a horizon (bounded by an ergotoroid rather than an ergosphere), the effective potentials look like Fig. 11.15, drawn for $aL < 0$ (Comins and Schutz 1978). (For $aL > 0$ the curves just turn over, of course.)

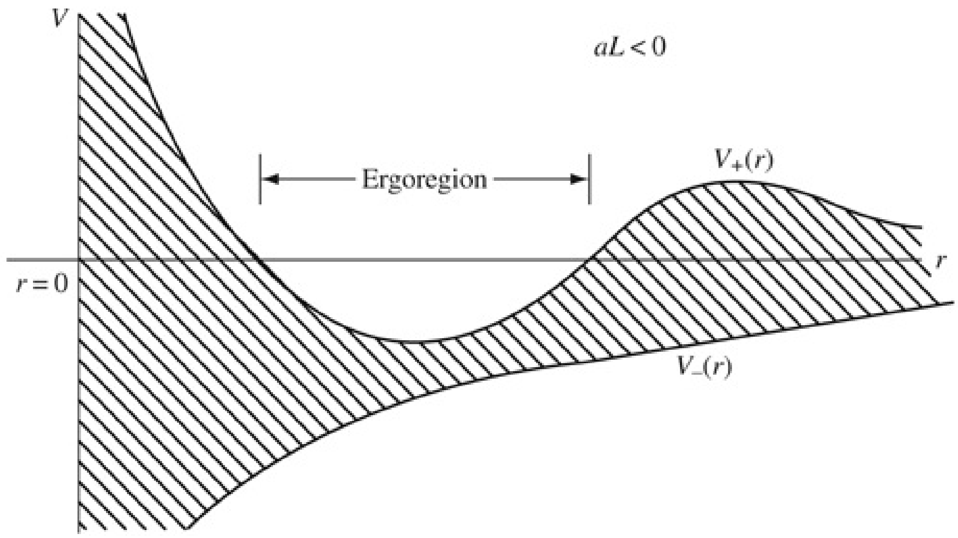


Figure 11.15 As Fig. 11.14 for equatorial orbits in the spacetime of a star rotating rapidly enough to have an ergoregion.

The curve for V_+ dips below zero in the ergoregion. Outside the ergoregion it is positive, climbing to infinity as $r \to 0$. The curve for V_- never changes sign, and also goes to infinity as $r \to 0$. The Penrose process operates inside the ergoregion, where the negative-energy photons are trapped.

In stars, the existence of an ergoregion leads to an instability (Friedman 1978). Roughly speaking, here is how it works. Imagine a small 'seed' gravitational wave that has, perhaps by scattering from the star, begun to travel on a negative-energy null line in the ergoregion. It is trapped there, and must circle around the star forever within the ergotoroid. But, being a wave, it is not localized perfectly, and so part of the wave inevitably creates disturbances outside the ergoregion, *i.e.* with positive energy, which radiate to infinity. By conservation of energy, the wave in the ergoregion must get stronger because of this 'leakage' of energy, since strengthening it means creating more negative energy to balance the positive energy radiated away. But then the cycle continues: the stronger wave in the ergoregion leaks even more energy to infinity, and so its amplitude grows even more. This is the hallmark of an exponential process: the stronger it gets, the faster it grows. Of course, these waves carry angular momentum away as well, so in the long run the process results in the spindown of the star and the complete disappearance of the ergoregion.

In principle, this might also happen to Kerr black holes, but here there is a crucial difference: the wave in the ergoregion does not have to stay there forever, but can instead travel across the horizon into the hole. There it can no longer create waves outside the ergosphere. Physicists have studied the stability of the Kerr metric intensively, and all indications are that all waves in the ergoregion lose more amplitude across the horizon than they gain by radiating to infinity: in other words, Kerr is stable. There are still gaps in the proof, but (as we shall see below) astronomical observations increasingly suggest that there are Kerr black holes in Nature, so the gaps will one day probably be filled by more clever mathematics than we have so far been able to bring to this problem!

11.4 Real black holes in astronomy

As we noted before, the unique stationary (time-independent) solution for a black hole is the Kerr metric. This extraordinary result makes studying black holes in the real world much simpler than we might have expected. Most real-world systems are so complicated that they can only be studied through idealizations; each star, for example, is individual, and astronomers work very hard to classify them, discover patterns in their appearance that give clues to their interiors, and generally build models that are complex enough to capture their important properties. But these models are still simplifications, since with 10^{57} particles a star has an enormous number of physical degrees of freedom. Not so with black holes. Provided they are stationary, they have just *two* intrinsic degrees of freedom: their mass and spin. In his 1975 Ryerson Lecture at the University of Chicago, long before the majority of astronomers had accepted that black holes were commonplace in the universe, the astrophysicist S. Chandrasekhar expressed his reaction to the uniqueness theorem in this way (Chandrasekhar 1987):

> In my entire scientific life, extending over forty-five years, the most shattering experience has been the realization that an exact solution of Einstein's equations of general relativity, discovered by the New Zealand mathematician, Roy Kerr, provides the absolutely exact representation of untold numbers of massive black holes that populate the universe. This shuddering before the beautiful, this incredible fact that a discovery motivated by a search after the beautiful in mathematics should find its exact replica in Nature, persuades me to say that beauty is that to which the human mind responds at its deepest and most profound.

The simplicity of the black hole model has made it possible to identify systems containing black holes based only on indirect evidence, on their effects on nearby gas and stars. Until gravitational wave detectors become sufficiently sensitive to detect radiation from black holes, this will be the only way to find them.

The small size of black holes means that, in order to gain reasonable confidence in such an identification, we have normally to make an observation either with very high angular resolution to see matter near the horizon or by using photons of very high energy, which originate from strongly compressed and heated matter near the horizon. Using orbiting X-ray observatories, astronomers since the 1970s have been able to identify a number of black hole candidates, particularly in binary systems in our Galaxy. These have masses around $10M_{\odot}$, and they presumably have arisen from the gravitational collapse of a massive star, as discussed in the previous chapter. With improvements in high-resolution astronomy, astronomers since the 1990s have been able to identify supermassive black holes in the centers of galaxies, with masses ranging from 10^6 to $10^{10}M_{\odot}$. It came as a big surprise when astronomers found that almost all galaxies that are close enough to allow the identification of a central black hole have turned out to contain one. Indeed, one of the most secure black hole identifications is the $4.3 \times 10^6 M_{\odot}$ black hole in the center of our own Galaxy (Genzel and Karas 2007)!

Black holes of stellar mass

An isolated black hole, formed by the collapse of a massive star, would be very difficult to identify. It might accrete a small amount of gas as it moves through the interstellar medium, but this gas would not emit much X-radiation before being swallowed. No such candidates have been identified. All known stellar-mass black holes are in binary systems whose companion star is so large that it begins dumping gas on to the hole. Being in a binary system, the gas has angular momentum, and so it forms a disk around the black hole. But within this disk there is friction, possibly caused by turbulence or by magnetic fields. Friction leads material to spiral inwards through the disk, giving up angular momentum and energy. Some of this energy goes into the internal energy of the gas, heating it up to temperatures in excess of 10^6 K, so that the peak of its emission spectrum is at X-ray wavelengths.

Many such X-ray binary systems are now known. Not all of them contain black holes: a neutron star is compact enough so that gas accreting on to it will also reach X-ray temperatures. Astronomers distinguish black holes from neutron stars in these systems by two means: mass and pulsations. If the accreting object pulsates in X-rays at a very steady rate, then it is a pulsar and it cannot be a black hole: black holes cannot hold on to a magnetic field and make it rotate with the hole's rotation. Most systems do not show such pulsations, however. In these cases, astronomers try to estimate the mass of the accreting object from observations of the velocity and orbital radius of the companion star (obtained by monitoring the Doppler shift of its spectral lines) and from an estimate of the companion's mass (again from its spectrum). These estimates are uncertain, particularly because it is usually hard to estimate the inclination of the plane of the orbit to the line of sight, but if the accreting object has an estimated mass much more than about $5M_{\odot}$, then it is believed likely to be a black hole. This is because the maximum mass of a neutron star cannot be much more than $3M_{\odot}$ and is likely much smaller.

Mass turns out to be a good discriminator: of the dozen or so black-hole candidates, only one or two have an estimated mass under $7M_{\odot}$, while all the mass estimates for the remaining X-ray binary systems lie around or under $2M_{\odot}$. There does seem to be a mass gap, therefore, between black holes and neutron stars, at least for objects formed in binaries. One of the most secure candidates is one of the first ever observed: Cyg X-1, whose mass is around $10M_{\odot}$. The largest stellar black-hole mass ever estimated is $70M_{\odot}$, for the black hole in the binary system M33 X-7. As its name indicates, this system lies in the galaxy M33, which at 1 Mpc is about twice as far from our Galaxy as the Andromeda Galaxy (M31) is. What is most remarkable about M33 X-7 is that it is an eclipsing system, which constrains the inclination angle enough to make the mass estimate more secure.

Although we are confident that stellar evolution occasionally produces black holes, the pathway is still in doubt. Most astronomers assume that this happens in supernova explosions, and this view is reinforced by the fact that computer simulations of supernovae have more difficulty expelling the stellar envelope and leaving a neutron star behind than they do forming a black hole from the collapsing stellar core. But there is increasing evidence to associate most gamma-ray bursts with black-hole formations from very rapidly rotating progenitor stars.

How do we know that these massive objects are black holes? The answer is that any other explanation is less plausible. They cannot be neutron stars: no equation of state that is causal (i.e. has a sound speed less than the speed of light) can support more than about $3M_{\odot}$. It would be possible to invent some kind of exotic matter (sometimes called bosonic matter) that might just make a massive compact object that does not collapse, but we have no evidence for such matter.

But until we detect the signature of black holes in gravitational waves, such as their ringdown radiation (Ch. 9), we will not be able to exclude such exotic scenarios completely.

Supermassive black holes

The best evidence for supermassive black holes is for the closest one to us: the $4.3 \times 10^6 M_\odot$ black hole in the center of the Milky Way. This was discovered by making repeated high-resolution measurements of the positions of stars in the very center of the Galaxy. The measurements had to be made using infrared light, because in visible light the center is obscured by interstellar dust. Over a period of ten years some of the stars were observed to move by very considerable distances, and not just on straight lines, but rather on clearly elliptical orbits. All the orbits had a common gravitating center, but the center was dark. The stars' spectra revealed their radial velocities, so with three-dimensional velocities and a good idea of how far away the galactic center is, it is possible to estimate the mass of the central object from each orbit. All the orbits are consistent, and point to a dark object of $4.3 \times 10^6 M_\odot$ directly at the center.

Remarkably, one of the orbiting stars approaches to within about 120 AU of the gravitating center, and attains a speed of some 5000 km s^{-1}, more than 1% of the speed of light! Other than a black hole, no known or plausible matter system could contain such a large amount of mass in such a small volume, without itself collapsing rapidly to a black hole.

Other black holes of similar masses are inferred in external galaxies, including in our nearest neighbor M31, mentioned above. It seems, therefore, that black holes are associated with galaxy formation, but the nature of this association is not clear: did the holes come first, or did they form as part of the process of galaxy formation? It is also not clear whether the holes formed with large mass, say 10^5–$10^6 M_\odot$, or whether they started out with smaller masses (perhaps 10^4–$10^5 M_\odot$) and grew later. And if they grew, it is not clear whether they grew by accreting gas and stars or by merging with other black holes. This last question may be answered by the LISA satellite (Ch. 9), which will detect mergers over a wide mass range throughout the universe.

But black holes like those in the center of our Galaxy are the babies of the supermassive black hole population: as we have mentioned earlier, astronomers believe that the quasar phenomenon is created by gas accreting on to much more massive black holes, typically $10^9 M_\odot$. While these are too far away for astronomers to resolve the region near the black hole, the only quasar model that has survived decades of observation in many wave bands is the black hole model. This model now has the consensus of the great majority of astrophysicists working in the field.

Since quasars were much more plentiful in the early universe than they are today, it seems that these ultra-massive black holes had to form very early, while their more modest counterparts like that, in the Milky Way might have taken longer. This suggests that the black holes in quasars did not form by the growth of holes like our own; this is another of the unanswered questions about supermassive black holes.

Remarkably, there seems to be a good relationship between the mass of the central black hole and the velocity dispersion (random velocities) in the central part of the galaxy surrounding the hole, which is called the galactic bulge. The more massive the hole, the higher the velocities. This relationship seems to have a simple form all the way from $10^6$to $10^{10} M_\odot$. It might be a clue to how the holes formed.

Astronomers know that galaxies frequently merge, and this ought to bring at least some of their black holes to merge as well, producing strong gravitational waves in the LISA band. In fact, as we shall see in the next chapter, current models for galaxy formation suggest that all galaxies are themselves the products of repeated mergers with smaller clusters of stars, and so it is possible that, in the course of the formation and growth of galaxies, the central black holes grew larger and larger by merging with incoming black holes.

As remarked before, LISA should decide this issue, but already there is a growing body of evidence for mergers of black holes. A number of galaxies with two distinct bright cores are known, and remarkable evidence was very recently (2008) announced for the ejection of a supermassive black hole at something like 1% of the speed of light from the center of a galaxy (Komossa *et al.* 2008). Speeds like this can only be achieved as a result of the ‘kick’ that a final black hole gets as a result of the merger (see below).

Intermediate-mass black holes

If there are black holes of $10 M_\odot$ and of $10^6 M_\odot$ and more, are there black holes with masses around $100–10^4 M_\odot$? Astronomers call these intermediate mass black holes, but the evidence for their existence is still ambiguous. Some bright

X-ray sources may be large black holes, and perhaps there are black holes in this mass range in globular clusters. Astronomers are searching intensively for better evidence for these objects, which might have formed in the first generation of star formation, when clouds made up of pure hydrogen and helium first collapsed. These first stars were probably a lot more massive than the stars, like our Sun, that formed later from gas clouds that had been polluted by the heavier elements made by the first generation of stars. Numerical simulations suggest that some of these first stars (called Population III stars) could have rapidly collapsed to black holes. But finding these holes will not be easy. As this book goes to press (2008), unpublished data from the European Space Agency suggest the existence of a black hole with mass $4 \times 10^4 M_\odot$ in the cluster Omega Centauri. Observations of similar objects may well reveal more such black holes.

Dynamical black holes

Although stationary black holes are simple, there are situations where black holes are expected to be highly dynamical, and these are more difficult to treat analytically. When a black hole is formed, any initial asymmetry (such as quadrupole moments) must be radiated away in gravitational waves, until finally only the mass and angular momentum are left behind. This generally happens quickly: studies of linear perturbations of black holes show that black holes have a characteristic spectrum of oscillations, but that they typically damp out (ring down) exponentially after only a few cycles (Kokkotas and Schmidt 1999). The Kerr metric takes over very quickly.

Even more dynamical are black holes in collision, either with other black holes or with stars. As described in Ch. 9, binary systems involving black holes will eventually merge, and black holes in the centers of galaxies can merge with other massive holes when galaxies merge. These situations can only be studied numerically, by using computers to solve Einstein's equations and perform a dynamical simulation.

Numerical techniques for GR have been developed over a period of several decades, but progress initially was slow. The coordinate freedom of general relativity, coupled with the complexity of the Einstein equations, means that there is no unique way to formulate a system of equations for the computer. Most formulations have turned out to lead to intrinsically unstable numerical schemes, and finding a stable scheme took much trial and error. Moreover, when black holes are involved the full metric has a singularity where its components diverge; this has somehow to be removed from the numerical domain, because computers can work only to a finite precision. See Bona and Palenzuela-Luque (2005) and Alcubierre (2008) for surveys of these problems and their solutions.

Only by the mid-2000s were scientists able to regulate all these problems and produce reliable simulations of spinning Kerr black holes orbiting one another and then merging together to form a single final Kerr black hole. Results have been coming out rapidly since then. One of the most interesting aspects of black hole mergers is the so-called ‘kick’. When there is no particular symmetry in the initial system, then the emitted gravitational radiation will emerge asymmetrically, so that the waves will carry away a net linear momentum in some direction. The result will be that the final black hole recoils in the opposite direction. The velocity of the recoil, being dimensionless, does not depend on the overall mass scale of the system, just on dimensionless initial data: the ratio of the masses of the initial holes, the dimensionless spin parameters of the holes (a/M), and the directions of the spins. Normally these recoil velocities are of order a few hundred km s^{-1}, which could be enough to expel the black hole from the center of a star cluster or even a spiral galaxy. Even more remarkably, recoil velocities exceeding 10% of the speed of light are inferred from simulations for some coalescences (Campanelli *et al.* 2007). In the long run, the predictions of gravitational wave emission from these simulations will be used by gravitational wave astronomers to assist in searches and in the interpretation of signals that are found.

11.5 Quantum mechanical emission of radiation by black holes: the Hawking process

In 1974 Stephen Hawking startled the physics community by proving that black holes aren’t black: they radiate energy continuously! This doesn’t come from any mistake in what we have already done; it arises in the application of quantum mechanics to electromagnetic fields near a black hole. We have until now spoken of photons as particles following a geodesic trajectory in spacetime; but according to the uncertainty principle these ‘particles’ cannot be localized to arbitrary precision. Near the horizon this markedly changes the behavior of ‘real’ photons from what we have already described for idealized null particles.

Hawking’s calculation (Hawking 1975) uses the techniques of quantum field theory, but we can derive its main prediction very simply from elementary considerations. What follows, therefore, is a ‘plausibility argument’, not a rigorous discussion of the effect. One form of the uncertainty principle is $\Delta E\ \Delta t$

$\geq \hbar/2$, where ΔE is the minimum uncertainty in a particle's energy which resides in a quantum mechanical state for a time Δt. According to quantum field theory, ordinary space is filled with 'vacuum fluctuations' in electromagnetic fields, which consist of pairs of photons being produced at one event and recombining at another. Such pairs violate conservation of energy, but if they last less than $\Delta t = \hbar/2\Delta E$, where ΔE is the amount of violation, they violate no physical law. Thus, in the largescale, energy conservation holds rigorously, while, on a small scale, it is always being violated. Now, as we have emphasized before, spacetime near the horizon of a black hole is perfectly ordinary and, in particular, locally flat. Therefore these fluctuations will also be happening there. Consider a fluctuation which produces two photons, one of energy E and the other with energy $-E$. In flat spacetime the negative-energy photon would not be able to propagate freely, so it would necessarily recombine with the positive-energy one within a time $\hbar/2E$. But if produced just outside the horizon, it has a chance of crossing the horizon before the time $\hbar/2E$ elapses; once inside the horizon it *can* propagate freely, as we shall now show. Consider the Schwarzschild metric for simplicity, and recall from our discussion of orbits in the Kerr metric that negative energy is normally excluded because it corresponds to a particle that propagates backwards in time. Inside the event horizon, an observer going forwards in time is one going toward decreasing r. For simplicity let us choose one on a trajectory for which $p_0 = 0 = U^0 = 0$. Then U^r is the only nonzero component of $\vec{U}$, and by the normalization condition $\vec{U} \cdot \vec{U} = -1$ we find U^r:

$$U^r = -\left(\frac{2M}{r} - 1\right)^{1/2}, \qquad r < 2M, \tag{11.98}$$

negative because the observer is ingoing. Any photon orbit is allowed for which $-\vec{p} \cdot \vec{U} > 0$. Consider a zero angular-momentum photon, moving radially inside the horizon. By Eq. (11.12) with $L = 0$, it clearly has $E = \pm\, p^r$. Then its energy relative to the observer is

$$-\vec{p} \cdot \vec{U} = -p^r U^r g_{rr} = -\left(\frac{2M}{r} - 1\right)^{-1/2} p^r. \tag{11.99}$$

This is positive if and only if the photon is also ingoing: $p^r < 0$. But it sets *no* restriction at all on E. Photons may travel on null geodesics inside the horizon, which have either sign of E, as long as $p^r < 0$. (Recall that t is a spatial coordinate inside the horizon, so this result should not be surprising: E is a spatial momentum component there.)

Since a fluctuation near the horizon *can* put the negative-energy photon into a realizeable trajectory, the positive-energy photon is allowed to escape to infinity. Let us see what we can say about its energy. We first look at the fluctuations in a freely falling inertial frame, which is the one for which spacetime is locally flat and in which the fluctuations should look normal. A frame that is momentarily at rest at coordinate $2M + \varepsilon$ will immediately begin falling inwards, following the trajectory of a particle with $\tilde{L} = 0$ and $\tilde{E} = [1 - 2M/(2M + \varepsilon)]^{1/2} \approx (\varepsilon/2M)^{1/2}$, from Eq. (11.11). It reaches the horizon after a proper-time lapse $\Delta\tau$ obtained by integrating Eq. (11.59):

$$\Delta\tau = -\int_{2M+\varepsilon}^{2M} \left(\frac{2M}{r} - \frac{2M}{2M+\varepsilon}\right)^{-1/2} \mathrm{d}r. \tag{11.100}$$

To first order in ε this is

$$\Delta\tau = 2(2M\varepsilon)^{1/2}. \tag{11.101}$$

We can find the energy $\mathcal{E}$ of the photon in this frame by setting this equal to the fluctuation time $\hbar/2\mathcal{E}$. The result is

$$\mathcal{E} = \tfrac{1}{4}\hbar(2M\varepsilon)^{-1/2}. \tag{11.102}$$

This is the energy of the outgoing photon, the one which reaches infinity, as calculated on the local inertial frame. To find its energy when it gets to infinity we recall that

$$\mathcal{E} = -\vec{p}\cdot\vec{U},$$

with $-U_0 = \tilde{E} \approx (\varepsilon/2M)^{1/2}$. Therefore

$$\mathcal{E} = -g^{00}p_0U_0 = U_0g^{00}E, \tag{11.103}$$

where E is the conserved energy on the photon's trajectory, and is the energy it is measured to have when it arrives at infinity. Evaluating g^{00} at $2M + \varepsilon$ gives, finally,

$$E = \mathcal{E}(\varepsilon/2M)^{1/2} = \hbar/8M. \tag{11.104}$$

Remarkably, it doesn't matter where the photon originated: it always comes out with this characteristic energy!

The rigorous calculation which Hawking performed showed that the photons which come out have the spectrum characteristic of a black body with a temperature

$$T_H = \hbar/8\pi kM, \tag{11.105}$$

where k is Boltzmann's constant. At the peak of the black-body spectrum, the energy of the photon is (by Wien's displacement law)

$$E = 4.965kT = 1.580\hbar/8M, \tag{11.106}$$

fairly close to our crude result, Eq. (11.104). Our argument does not show that the photons should have a black-body spectrum; but the fact that the spectrum originates in random fluctuations, plus the fact that the black hole is, classically, a perfect absorber, makes this result plausible as well.

It is important to understand that the negative-energy photons in the Hawking effect are not the same as the negative-energy photons that we discussed in the Penrose process above. The Penrose process works only inside an ergoregion, and uses negative-energy orbits that are outside the horizon of the black hole. The Hawking result is more profound: it operates even for a nonspinning black hole and connects negative-energy photons *inside* the horizon with positive-energy counterparts outside. It operates in the Kerr metric as well, but again it happens across the horizon, not the ergosphere. The Hawking effect does not lead to an unstable runaway, the way the Penrose process does for a star with an ergoregion. This is because Hawking's negative-energy photon is already inside the horizon and does not create any further positive-energy photons outside. So the Hawking radiation is a steady thermal radiation, created by ever-present quantum fluctuations near the horizon.

Notice that the Hawking temperature of the hole is proportional to M^{-1}. The rate of radiation from a black body is proportional to AT^4, where A is the area of the body, in this case of the horizon, which is proportional to M^2 (see Eq. (11.85)). So the luminosity of the hole is proportional to M^{-2}. This energy must come from the mass of the hole (every negative-energy photon falling into it decreases M), so we have

$$\left.\begin{aligned} dM/dt &\sim M^{-2}, \\ M^2\, dM &\sim dt, \end{aligned}\right\} \tag{11.107}$$

or the lifetime of the hole is

$$\tau \sim M^3. \tag{11.108}$$

The bigger the hole the longer it lives, and the cooler its temperature. The numbers work out that a hole of mass 10^{12} kg has a lifetime of 10^{10} yr, about the age of the universe. Thus

$$\left(\frac{\tau}{10^{10}\,\text{yr}}\right) = \left(\frac{M}{10^{12}\,\text{kg}}\right)^3. \tag{11.109}$$

Since a solar mass is about 10^{30} kg, black holes formed from stellar collapse are essentially unaffected by this radiation, which has a temperature of about 10^{-7} K. On the other hand, it is possible for holes of 10^{12} kg to form in the very early universe. To see the observable effect of their 'evaporation', let us calculate the energy radiated in the last second by setting $\tau = 1\ \text{s} = (3 \times 10^7)^{-1}$ yr in Eq. (11.109). We get

$$M \approx 10^6\,\text{kg} \sim 10^{23}\,\text{J}. \tag{11.110}$$

So for a brief second it would have a luminosity about 0.1% of the Sun's luminosity, but in spectrum it would be very different. Its temperature would be 10^{11} K, emitting primarily in γ-rays! We might be tempted to explain the gamma-ray bursts mentioned earlier in this chapter as primordial black hole evaporations, but the observed gamma bursts are in fact billions of times more luminous. A primordial black-hole evaporation would probably be visible only if it happened in our own Galaxy. No such events have been identified.

It must be pointed out that all derivations of Hawking's result are valid only if the typical photon has $E \ll M$, since they involve treating the spacetime of the black hole as a fixed background in which we solve the equations of quantum mechanics, unaffected to first order by the propagation of these photons. This approximation fails for $M \approx h/M$, or for black holes of mass

$$M_{\text{Pl}} = h^{1/2} = 1.6 \times 10^{-35}\ \text{m} = 2.2 \times 10^{-8}\ \text{kg}. \tag{11.111}$$

This is called the Planck mass, since it is a mass derived only from Planck's constant (and c and G). To treat quantum effects involving such holes, we need a consistent theory of quantum gravity, which is one of the most active areas of research in relativity today. All we can say here is that the search has not yet proved fully successful, but Hawking's calculation appears to have been one of the most fruitful steps.

The Hawking effect has provided a remarkable unification of gravity and thermodynamics. Consider Hawking's area theorem, which we may write as

$$\frac{\mathrm{d}A}{\mathrm{d}t} \geqslant 0. \tag{11.112}$$

For a Schwarzschild black hole,

$$A = 16\,\pi M^2,$$
$$\mathrm{d}A = 32\,\pi M\,\mathrm{d}M,$$

or (if we arrange factors in an appropriate way)

$$\mathrm{d}M = \frac{1}{32\,\pi M}\,\mathrm{d}A = \frac{\hbar}{8\,\pi kM}\,\mathrm{d}\left(\frac{kA}{4\hbar}\right). \tag{11.113}$$

Since $\mathrm{d}M$ is the change in the hole's total energy, and since $\hbar/8\pi kM$ is its Hawking temperature T_H, we may write Eq. (11.113) in the form

$$\mathrm{d}E = T_H\,\mathrm{d}S,$$

with

$$S = kA/4\hbar. \tag{11.114}$$

Since, by Eq. (11.112), this quantity S can never decrease, we have in Eqs. (11.113) and (11.112) the first and second laws of thermodynamics as they apply to black holes! That is, a black hole behaves in every respect as a thermodynamic black body with temperature $T_H = \hbar/8\pi kM$ and entropy $kA/4\hbar$. This analogy had been noticed as soon as the area theorem was discovered (see Bekenstein 1973, 1974, and Misner *et al.* 1973, Box 33.4), but at that time it was thought to be an incomplete analogy because black holes did not have a true temperature. The Hawking radiation fits the missing piece into the puzzle.

But the Hawking radiation has raised other questions. One of them concerns information. The emission of radiation from the black hole raises the possibility that the radiation could carry information that, in the classical picture, is effaced by the formation of the horizon. If the radiation is perfectly thermal, then it contains no information. But it is possible that the outgoing photons and gravitons have a thermal black-body spectrum but also have weak correlations that contain the information. This information would not have come from inside the hole, but from the virtual photons and gravitons just outside the horizon, which are affected by the details of the collapsing matter that passes through them on its the way to forming the black hole, and which then become real photons and gravitons by the process described above. Whether this picture is indeed correct, and what kind of information can in principle be recovered from the outgoing radiation, are still matters of considerable debate among physicists.

The Hawking radiation has also become a touchstone for the development of full theories of quantum gravity. Physicists test new quantization methods by showing that they can predict the Hawking radiation and associated physics, such as the entropy of the black hole. This is not sufficient to prove that a method will work in general, but it is regarded as necessary.

11.6 Further reading

The story of Karl Schwarzschild's discovery of the solution named after him is an extraordinary one. See the on-line biography of Schwarzschild by J. J. O'Connor and E. F. Robertson in the MacTutor History of Mathematics archive, at the URL `www-history.mcs.st-andrews.ac.uk/Biographies/Schwarzschild.html` (cited April 2008).

The perihelion shift and deflection of light are the two classical tests of GR. Other theories predict different results: see Will (1993, 2006). A short, entertaining account of the observation of the deflection of light and its impact on Einstein's fame is in McCrea (1979). To learn more about gravitational lensing, see Wambsganss (1998), Schneider *et al* (1992), Schneider (2006), or Perlick (2004).

The Kerr metric has less symmetry than the Schwarzschild metric, so it might be expected that particle orbits would have fewer conserved quantities and therefore be harder to calculate. This is, quite remarkably, false: even orbits out of the equator have three conserved quantities: energy, angular momentum, and a difficult-to-interpret quantity associated with the θ motion. The same remarkable property carries over to the wave equations that govern electromagnetic fields and gravitational waves in the Kerr metric: these equations separate completely in certain coordinate systems. See Teukolsky (1972) for the first general proof of this and Chandrasekhar (1983) for full discussions.

Black-hole thermodynamics is treated thoroughly in Carter (1979), while the related theory of quantum fields in curved spacetimes is reviewed by Wald (1994). This relates to work on quantum gravity. See the references cited in the §7.5.

Although the field of numerical relativity is rapidly evolving, some parts of it are mature and have been reviewed. Interested readers might consult the numerical relativity section of the online journal *Living Reviews in Relativity* at `http://relativity.livingreviews.org/` or, at a more popular level, the Einstein Online website `http://www.einstein-online.info/en/`.

11.7 Exercises

Consider a particle or photon in an orbit in the Schwarzschild metric with a certain E and L, at a radius $r \gg M$. Show that if spacetime were really *flat*, the particle would travel on a straight line which would pass a distance $b := L/[E^2 - m^2]^{1/2}$ from the center of coordinates $r = 0$. This ratio b is called the *impact parameter*. Show also that photon orbits that follow from Eq. (11.12) depend *only* on b.

Prove Eqs. (11.17) and (11.18).

Plot $\tilde{V}^2$ against r/M for the three cases $\tilde{L}^2 = 25M^2$, $\tilde{L}^2 = 12M^2$, $\tilde{L}^2 = 9M^2$ and verify the qualitative correctness of Figs. 11.1 and 11.3.

What kind of orbits are possible outside a star of radius (a) 2.5 M, (b) 4 M, (c) 10 M?

The centers of active galaxies and quasars contain black holes of mass $10^6 - 10^9$ $M_\odot$ or more.

(a) Find the radius $R_{0.01}$ at which $-g_{00}$ differs from the 'Newtonian' value $1 - 2\,M/R$ by only 1%. (We may think of this as a kind of limit on the region in which relativistic effects are important.)

(b) A 'normal' star may have a radius of 10^{10} m. Approximately how many such stars could occupy the volume of space between the horizon $R = 2\,M$ and $R_{0.01}$?

Compute the wavelength of light that gets to a distant observer from the following sources.

(a) Light emitted with wavelength 6563 Å (*Hα* line) by a source at rest where $\Phi = -10^{-6}$. (Typical star.)

(b) Same as (a) for $\Phi = -6 \times 10^{-5}$ (value for the white dwarf 40 Eridani B).

(c) Same as (a) for a source at rest at radius $r = 2.2\,M$ outside a black hole of mass $M = 1\,M_{\odot} = 1.47 \times 10^{5}$ cm.

(d) Same as (c) for $r = 2.02\,M$.

A clock is in a circular orbit at $r = 10\,M$ in a Schwarzschild metric.

(a) How much time elapses on the clock during one orbit? (Integrate the proper time $d\tau = |ds^2|^{1/2}$ over an orbit.)

(b) It sends out a signal to a distant observer once each orbit. What time interval does the distant observer measure between receiving any two signals?

(c) A second clock is located at rest at $r = 10\,M$ next to the orbit of the first clock. (Rockets keep it there.) How much time elapses on it between successive passes of the orbiting clock?

(d) Calculate (b) again in seconds for an orbit at $r = 6\,M$ where $M = 14\,M_{\odot}$. This is the minimum fluctuation time we expect in the X-ray spectrum of Cyg X-1: why?

(e) If the orbiting 'clock' is the twin Artemis, in the orbit in (d), how much does she age during the time her twin Diana lives 40 years far from the black hole and at rest with respect to it?

(a) Derive Eqs. (11.20) and (11.24).

(b) Derive Eqs. (11.26) and (11.28).

(This problem requires access to a computer.)

(a) Integrate numerically Eq. (11.26) or Eq. (11.28) for the orbit of a particle (i.e. for r/M as a function of ϕ) when $E^2 = 0.91$ and $(\tilde{L}/M)^2 = 13.0$. Compare the perihelion shift from one orbit to the next with Eq. (11.37).

(b) Integrate again when $\tilde{E}^2 = 0.95$ and $(\tilde{L}/M)^2 = 13.0$. How much proper time does this particle require to reach the horizon from $r = 10\,M$ if its initial radial velocity is negative?

) For a given value of $\tilde{L}$, what is the minimum value of $\tilde{E}$ that permits a particle with $m \neq 0$ to reach the Schwarzschild horizon?

(b) Express this result in terms of the *impact parameter b* (see Exer. 1).

(c) Conversely, for a given value of b, what is the maximum value of $\tilde{L}$ that permits a particle wth $m \neq 0$ to reach the Schwarzschild horizon? Relate your result to Fig. 11.3.

The right-hand side of Eq. (11.28) is a polynomial in u. Trace the u^3 term back through the derivation and show that it would not be present if we had started with the Newtonian version of Eq. (11.9). Interpret this term as a redshift effect on the orbital kinetic energy. Show that it is responsible for the maximum in the curve in Fig. 11.1.

(a) Prove that Eq. (11.32) solves Eq. (11.31).

(b) Derive Eq. (11.33) from Eq. (11.32) and show that it describes an ellipse by transforming to Cartesian coordinates.

(a) Derive Eq. (11.34) in the approximation that y is small. What must it be small compared to?

(b) Derive Eqs. (11.35) and (11.36) from (11.34).

(c) Verify the remark after Eq. (11.36) that $y = 0$ is not the correct circular orbit for the given $\tilde{E}$ and $\tilde{L}$ by using Eqs. (11.20) and (11.21) to find the correct value of y and comparing it to y_0 in Eq. (11.36).

(d) Show from Eq. (11.13) that a particle which has an inner turning point in the 'Newtonian' regime, *i.e.* for $r \gg M$, has a value $\tilde{L} \gg M$. Use this to justify the step from Eq. (11.37) to Eq. (11.38).

Compute the perihelion shift per orbit and per year for the following planets, given their distance from the Sun and their orbital period: Venus (1.1×10^{11} m, 1.9×10^7 s); Earth (1.5×10^{11} m, 3.2×10^7 s); Mars (2.3×10^{11} m, 5.9×10^7 s).

a) Derive Eq. (11.51) from (11.49), and show that it describes a straight line passing a distance b from the origin.

(b) Derive Eq. (11.53) from (11.49).

(c) Integrate Eq. (11.53) to get (11.54).

We calculate the observed deflection of a null geodesic anywhere on its path as follows. See Ward (1970).

(a) Show that Eq. (11.54) may be solved to give

$$bu = \sin(\phi - \phi_0) + \frac{M}{b}[1 - \cos(\phi - \phi_0)]^2 + 0\left(\frac{M^2}{b^2}\right). \tag{11.115}$$

(b) In Schwarzschild coordinates, the vector

$$\vec{v} \to -(0, 1, 0, \mathrm{d}\phi/\mathrm{d}r) \tag{11.116}$$

is tangent to the photon's path as seen by an observer at rest in the metric at the position r. Show that this observer measures the angle α in Fig. 11.16 to be

$$\cos\alpha = (\vec{v}\cdot\vec{e}_r)/(\vec{v}\cdot\vec{v})^{1/2}(\vec{e}_r\cdot\vec{e}_r)^{1/2}, \tag{11.117}$$

where $\vec{e}_r$ has components (0, 1, 0, 0). Argue that $\phi - \pi + \alpha$ is the *apparent* angular position of the star, and show from Eq. (11.115) that if $M = 0$ (no deflection), $\phi - \pi + \alpha = \phi_0$.

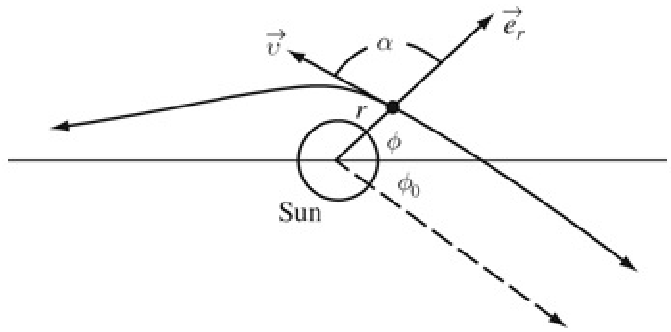


Figure 11.16 The deflection of light by the Sun.

(c) When $M \neq 0$, calculate the deflection

$$\delta\phi := (\phi - \pi + \alpha) - \phi_0 \tag{11.118}$$

to first order in M/b. Don't forget to use the Schwarzschild metric to compute the dot products in Eq. (11.117). Obtain

$$\delta\phi = \frac{2M}{b}[1 - \cos(\phi - \phi_0)], \tag{11.119}$$

which is, in terms of the position r of the observer,

$$\delta\phi = \frac{2M}{r}\frac{1 - \cos(\phi - \phi_0)}{\sin(\phi - \phi_0)}. \tag{11.120}$$

(d) For M=1 M =1.47 km, r=1 AU=1.5×10^6 km, how far from the Sun on the sky can this deflection be detected if we can measure angles to an accuracy of 2×10^{-3}arcsec?

We can use Eq. (11.115) above on a different problem, namely to calculate the expected arrival times at a distant observer of pulses regularly emitted by a satellite in a circular orbit in the Schwarzschild metric. This is a simplified version of the timing problem for the binary pulsar system (§ 9.4). See Damour and Deruelle (1986).

(a) Show that along the trajectory, Eq. (11.115), coordinate time elapses at the rate

$$dt/d\phi = b\left[(bu)^2\left(1 - \frac{2M}{b}bu\right)\right]^{-1}. \tag{11.121}$$

(b) Integrate this to find the coordinate travel time for a photon emitted at the position u_E, ϕ_E and received at the position u_R, ϕ_R, where $u_R \ll u_E$.

(c) Since Eq. (11.115) is satisfied at both (u_R, ϕ_R) and (u_E, ϕ_E), show that

$$\begin{aligned}\phi_R - \phi_0 = (u_R/u_E)\sin(\phi_E - \phi_R)\Big\{&1 + (u_R/u_E)\cos(\phi_E - \phi_R)\\ &+ Mu_E(1 - \cos[\phi_E - \phi_R])^2/\sin^2(\phi_E - \phi_R)\Big\},\end{aligned} \tag{11.122}$$

to first order in Mu_E and u_R/u_E and that, similarly,

$$b = (1/u_R)\left\{\phi_R - \phi_0 + Mu_E[1 - \cos(\phi_E - \phi_R)]^2/\sin(\phi_E - \phi_R)\right\}. \tag{11.123}$$

(d) Use these in your result in (b) to calculate the difference δt in travel time between pulses emitted at (u_E, ϕ_E) and at (u_E, $\phi_E + \delta\phi_E$) to first order in $\delta\phi_E$. (The receiver is at fixed (u_R, ϕ_R).)

(e) For an emitter in a circular orbit u_E = const., $\phi_E = \Omega t_E$, plot the relativistic corrections to the arrival time interval between successive pulses as a function of observer 'time', Ωt_R. Comment on the use of this graph, in view of the original assumption $M/b \ll 1$.

Use the expression for distances on a sphere, Eq. (10.2), to show that all the points on the line $\theta = 0$ in Fig. 11.9 are the same physical point.

Derive Eqs. (11.59) and (11.60).

a) Using the Schwarzschild metric, compute all the nonvanishing Christoffel symbols:

$$\Gamma^t{}_{rt} = -\Gamma^r{}_{rr} = \frac{M}{r^2}\left(1-\frac{2M}{r}\right)^{-1}; \quad \Gamma^r{}_{tt} = \frac{M}{r^2}\left(1-\frac{2M}{r}\right),$$
$$\Gamma^r{}_{\theta\theta} = \Gamma^r{}_{\phi\phi}/\sin^2\theta = -r\left(1-\frac{2M}{r}\right),$$
$$\Gamma^\theta{}_{\theta r} = \Gamma^\phi{}_{\phi r} = \frac{1}{r},$$
$$\Gamma^\phi{}_{\theta\phi} = -\Gamma^\theta{}_{\phi\phi}/\sin^2\theta = \cot\theta. \tag{11.124}$$

Show that all others vanish or are obtained from these by symmetry. (In your argument that some vanish, you should use the symmetries $t \rightarrow -t$, $\phi \rightarrow -\phi$, under either of which the metric is invariant.)

(b) Use (a) or the result of Exer. 35, § 6.9 to show that the only nonvanishing components of the Riemann tensor are

$$\begin{aligned} R^t{}_{rtr} &= -2\frac{M}{r^3}\left(1-\frac{2M}{r}\right)^{-1}, \\ R^t{}_{\theta t\theta} &= R^t{}_{\phi t\phi}/\sin^2\theta = M/r, \\ R^\theta{}_{\phi\theta\phi} &= 2M\sin^2\theta/r, \\ R^r{}_{\theta r\theta} &= R^r{}_{\phi r\phi}/\sin^2\theta = -M/r, \end{aligned} \tag{11.125}$$

plus those obtained by symmetries of the Riemann tensor.

(c) Convert these components to an *orthonormal* basis aligned with the Schwarzschild coordinates. Show that all components fall off as r^{-3} for large r.

(d) Compute $R^{\alpha\beta\mu\nu}R_{\alpha\beta\mu\nu}$, which is independent of the basis, and show that it is singular as $r \rightarrow 0$.

A particle of $m \neq 0$ falls radially toward the horizon of a Schwarzschild black hole of mass M. The geodesic it follows has $\tilde{E} = 0.95$.

(a) Find the proper time required to reach $r = 2\,M$ from $r = 3\,M$.

(b) Find the proper time required to reach $r = 0$ from $r = 2\,M$.

(c) Find, on the Schwarzschild coordinate basis, its four-velocity components at $r = 2.001\,M$.

(d) As it passes $2.001\,M$, it sends a photon out radially to a distant stationary observer. Compute the redshift of the photon when it reaches the observer. Don't forget to allow for the Doppler part of the redshift caused by the particle's velocity.

A measure of the tidal force on a body is given by the equation of geodesic deviation, Eq. (6.87). If a human will be crushed when the acceleration gradient across its body is 400 m s^{-2} per meter, calculate the minimum mass Schwarzschild black hole that would permit a human to survive long enough to reach the horizon on the trajectory in Exer. 21.

Prove Eq. (11.67).

Show that spacetime is locally flat at the center of the Kruskel–Szekeres coordinate system, $u = v = 0$ in Fig. 11.11.

Given a spherical star of radius $R \gg M$ and mean density ρ, estimate the tidal force across it which would be required to break it up. Use this as in Exer. 22 to define the tidal radius R_T of a black hole of mass M_H: the radius at which a star of density ρ near the hole will be torn apart. For what mass M_H is $R_T = 100\ M_H$ if $\rho = 10^3$ kg m^{-3}, typical of our Sun? This illustrates that even some applications of black holes in astrophysical contexts require few 'relativistic' effects.

Given the area of a Kerr hole, Eq. (11.85), with r_+ defined in Eq. (11.81), show that any two holes with masses m_1 and m_2 and angular momenta m_1a_1 and m_2a_2 respectively have a total area less than that of a single hole of mass $m_1 + m_2$ and angular momentum $m_1a_1 + m_2a_2$.

Show that the 'static limit', Eq. (11.80), is a limit on the region of spacetime in which curves with r, θ, and ϕ constant are timelike.

) Prove Eq. (11.87).

(b) Derive Eq. (11.89).

In the Kerr metric, show (or argue on symmetry grounds) that a geodesic which passes through a point in the equatorial 'plane' ($\theta = \pi/2$) and whose tangent there is tangent to the plane ($p^\theta = 0$) remains always in the plane.

Derive Eqs. (11.91) and (11.92).

Show that a ZAMO has four-velocity components $U^0 = A$, $U^\phi = \omega A$, $U^r = U^\theta = 0$, $A^2 = g_{\phi\phi}/(-D)$, where D is defined in Eq. (11.87).

Show, as argued in the text, that the Penrose process decreases the angular momentum of the hole.

Derive Eq. (11.101) from Eq. (11.100).

a) Use the area theorem to calculate the *maximum* energy released when two Schwarzschild black holes of mass M collide to form a Schwarzschild hole.

(b) Do the same for holes of mass m_1 and m_2, and express the result as a percentage of m_1 when $m_1 \to 0$ for fixed m_2.

The Sun rotates with a period of approximately 25 days.

(a) Idealize it as a solid sphere rotating uniformly. Its moment of inertia is $\frac{2}{5}M_\odot R_\odot^2$, where $M_\odot = 2 \times 10^{30}$ kg and $R_\odot = 7 \times 10^8$ m. In SI units compute $J_\odot$.

(b) Convert this to geometrized units.

(c) If the entire Sun suddenly collapsed into a black hole, it would form a Kerr hole of mass $M_\odot$ and angular momentum $J_\odot$. What would be the Kerr parameter, $a_\odot = J_\odot/M_\odot$, in m? What is the ratio $a_\odot/M_\odot$? Physicists expect that a Kerr hole will *never* be formed with $a > M$, because centrifugal forces will halt the collapse or create a rotational instability. The result of this exercise is that even a quite ordinary star like the sun needs to get rid of angular momentum before forming a black hole.

(d) Does an electron have too much angular momentum to form a Kerr hole with $a < M$? (Neglect its charge.)

) For a Kerr black hole, prove that for fixed M, the largest area is obtained for $a = 0$ (Schwarzschild).

(b) Conversely, prove that for fixed area, the smallest mass is obtained for $a = 0$.

) An observer sits at constant r, ϕ in the equatorial plane of the Kerr metric ($\theta = \pi/2$) outside the ergoregion. He uses mirrors to cause a photon to circle the hole along a circular path of constant r in the equatorial plane. Its world line is thus a null line with $\mathrm{d}r = \mathrm{d}\theta = 0$, but it is not, of course, a geodesic. How much coordinate time t elapses between the emission of a photon in the direction of increasing ϕ and its receipt after it has circled the hole once? Answer the same for a photon sent off in the direction of decreasing ϕ, and show that this is a different amount of time. Does the photon return redshifted from its original frequency?

(b) A different observer rotates about the hole on an orbit of r = const. and angular velocity given by Eq. (11.77). Using the same arrangement of mirrors, he measures the coordinate time that elapses between his emission and his receipt of a photon sent in either direction. Show that in this case the two terms are *equal*. (This is a ZAMO, as defined in the text.)

Consider equatorial motion of particles with $m \neq 0$ in the Kerr metric. Find the analogs of Eqs. (11.91)–(11.95) using $\tilde{E}$ and $\tilde{L}$ as defined in Eqs. (11.5) and (11.6). Plot $\tilde{V}_\pm$ for $a = 0.5\ M$ and $\tilde{L}/M = 20$, 12, and 6. Discuss the qualitative features of the trajectories. For arbitrary a determine the relations among $\tilde{E}$, $\tilde{L}$, and r for circular orbits with either sense of rotation. What is the minimum radius of a stable circular orbit? What happens to circular orbits in the ergosphere?

) Derive Eq. (11.109) from Eq. (11.105) and the black-body law, luminosity = σAT^4, where A is the area and σ is the Stefan–Boltzmann radiation constant, $\sigma = 0.567 \times 10^{-7}\ \mathrm{Wm^{-2}(K)^{-4}}$.

(b) How small must a black hole be to be able to emit substantial numbers of electron– positron pairs?

12
Cosmology

12.1 What is cosmology ?

The universe in the large

Cosmology is the study of the universe as a whole: its history, evolution, composition, dynamics. The primary aim of research in cosmology is to understand the large-scale structure of the universe, but cosmology also provides the arena, and the starting point, for the development of all the detailed small-scale structure that arose as the universe expanded away from the Big Bang: galaxies, stars, planets, people. The interface between cosmology and other branches of astronomy, physics, and biology is therefore a rich area of scientific research. Moreover, as astronomers have begun to be able to study the evidence for the Big Bang in detail, cosmology has begun to address very fundamental questions of physics: what are the laws of physics at the very highest possible energies, how did the Big Bang happen, what came before the Big Bang, how did the building blocks of matter (electrons, protons, neutrons) get made? Ultimately, the origin of every system and structure in the natural world, and possibly even the origin of the physical laws that govern the natural world, can be traced back to some aspect of cosmology.

Our ability to understand the universe on large scales depends in an essential way on general relativity. It is not hard to see why. Newtonian theory is an adequate description of gravity as long as, roughly speaking, the mass M of a system is small compared to the size, $R : M/R \ll 1$. We must replace Newtonian theory with GR if the system changes in such a way that M/R gets close to one. This can happen if the system's radius R becomes small faster than M, which is the domain of compact or collapsed objects: neutron stars and black holes have very small radii for the mass they contain. But we can also get to the relativistic regime if the system's mass increases faster than its radius. This is the case for cosmology: if space is filled with matter of roughly the same density everywhere, then, as we consider volumes of larger and larger radius R, the mass increases as R^3, and M/R eventually must get so large that GR becomes important.

What length scale is this? Suppose we begin increasing R from the center of our Sun. The Sun is nowhere relativistic, and once R is larger than $R_{\odot}$, M hardly increases at all until the next star is reached. The system of stars of which the Sun is a minor member is a galaxy, and contains some 10^{11} stars in a radius of about 15 kpc. (One parsec, abbreviated pc, is about 3×10^{16} m.) For this system, $M/R \sim 10^{-6}$, similar to that for the Sun itself. So galactic dynamics has no need for relativity. (This applies to the galaxy as a whole: small regions, including the very center, may be dominated by black holes or other relativistic objects.) Galaxies are observed to form clusters, which often have thousands of members in a volume of the order of a Mpc. Such a cluster could have $M/R \sim 10^{-4}$, but it would still not need GR to describe it adequately.

When we go to larger scales than the size of a typical galaxy cluster, however, we enter the domain of *cosmology*.

In the cosmological picture, galaxies and even clusters are very small-scale structures, mere atoms in the larger universe. Our telescopes are capable of seeing to distances greater than 10 Gpc. On this large scale, the universe is observed to be *homogeneous*, to have roughly the same density of galaxies, and roughly the same types of galaxies, everywhere. As we shall see later, the mean density of mass–energy is roughly $\rho = 10^{-26}$ kg m^{-3}. Taking this density, the mass $M = 4\pi\rho R^3/3$ is equal to R for $R \sim 6$ Gpc, which is well within the observable universe. So to understand the universe that our telescopes reveal to us, we need GR.

Indeed, GR has provided scientists with their first consistent framework for studying cosmology. We shall see that metrics exist that describe universes that embody the observed homogeneity: they have no boundaries, no edges, and are homogeneous everywhere. Newtonian gravity could not consistently make such models, because the solution of Newton's fundamental equation $\nabla^2\Phi = 4\pi G\rho$ is ambiguous if there is no outer edge on which to set a boundary condition for the differential equation. So only with Einstein could cosmology become a branch of physics and astronomy.

We should ask the converse question: if we live in a universe whose overall structure is highly relativistic, how is it that we can study our local region of the universe without reference to cosmology? How can we, as in earlier chapters, apply general relativity to the study of neutron stars and black holes as if they were embedded in an empty asymptotically flat spacetime, when actually they exist in a highly relativistic cosmology? How can astronomers study individual stars, geologists individual planets, biologists individual cells – all without reference to GR? The answer, of course, is that in GR spacetime is locally flat: as long as your experiment is confined to the local region you don't need to know about the large-scale geometry. This separation of local and global is not possible in Newtonian gravity, where even the local gravitational field within a large uniform-density system depends on the boundary conditions far away, on the shape of the distant "edge" of the universe (see Exer. 3, § 12.6). So GR not only allows us to study cosmology, it explains why we can study the rest of science without needing GR!

The cosmological arena

In recent years, with the increasing power of ground-and space-based astronomical observatories, cosmology has become a precision science, one which physicists look to for answers to some of their most fundamental questions. The basic picture of the universe that observations reveal is remarkably simple, when averaged over distance scales much larger than, say, 10 Mpc. We see a homogeneous universe, expanding at the same rate everywhere. The universe we see is also *isotropic:* it looks the same, on average, in every direction we look. The universe is filled with radiation with a black-body thermal spectrum, with a temperature of 2.725K. The expansion means that the universe has a finite age, or at least that it has expanded in a finite time from a state of very high density. The thermal radiation suggests that the universe was initially much hotter than today, and has cooled as it expanded. The expansion resolves the oldest of all cosmological conundrums, Olbers' Paradox. The sky is dark at night because we do not receive light from all stars in our infinite homogeneous universe, but only from stars that are close enough for light to have traveled to us during the age of the universe.

But the expansion raises other deep questions, about how the universe evolved to its present state and what it was like much earlier. We would like to know how the first stars formed, why they group into galaxies, why galaxies form clusters: where did the density irregularities come from that have led to the enormously varied structure of the universe on scales smaller than 10 Mpc? We would like to know how the elements formed, what the universe was like when it was too dense and too hot to have normal nuclei, and what the very hot early universe can tell us about the laws of physics at energies higher than we can explore with particle accelerators. We would like to know if the observed homogeneity and isotropy of the universe has a physical explanation.

Answering these questions has led physicists to explore some very deep issues at the frontiers of our understanding of fundamental physics. The homogeneity problem can be solved if the extremely early universe expanded exponentially rapidly, in a phase that physicists call *inflation*. This could happen if the laws of physics at higher energies than can be explored in the laboratory have a suitable form, and if so this would as a bonus help to explain the density fluctuations that led to the observed galaxies and clusters. As we shall see, it appears that most of the matter in the universe is in an unknown form, which physicists call *dark matter* because it radiates no light. Even more strangely, the universe seems to be pervaded by a relativistic energy density that carries negative pressure and which is driving the expansion faster and faster; physicists call this the *dark energy*. The mysteries of dark energy and of inflation may only really be solved with a better understanding of the laws of physics at the highest energies, so theoretical physicists are looking more and more to astronomical answers for clues to better theories.

Modern cosmology is already providing answers to some of these questions, and the answers are becoming more precise and more definite at a rapid pace. This chapter gives a snapshot of the fundamentals of our understanding at the present time (2008). More than any other area covered in this textbook, cosmology is a study that promises new insights, surprises, perhaps even a revolution.

12.2 Cosmological kinematics: observing the expanding universe

Before we can begin to understand the deep questions of cosmology, let alone their answers, we need to be able to describe and work with the notion of an expanding universe. In this section we develop the metric that describes a homogeneous expanding universe, we show how astronomical observations measure the expansion history, and we develop the framework for discussing physical processes in the expanding universe. In the following section 12.3 we will apply Einstein's equations to our models to see what GR has to tell us about how the universe expands.

Homogeneity and isotropy of the universe

The simplest approach to applying GR to cosmology is to use the remarkable observed large-scale uniformity. We see, on scales much larger than 10 Mpc, not only a uniform average density but uniformity in other properties: types of galaxies, their clustering densities, their chemical composition and stellar composition. Of course, when we look very far away we are also looking back in time, the time it took the light we observe to reach us; over sufficiently long look-back times we also see evolution, we see a younger universe. But the evolution we see is again the same in all directions, even when we look at parts of the early universe that are very far from one another. We therefore conclude that, on the large scale, the universe is *homogeneous*. What is more, on scales much larger than 10 Mpc the universe seems to be *isotropic* about every point: we see no consistently defined special direction.[1]

A third feature of the observable universe is the uniformity of its expansion: galaxies, on average, seem to be receding from us at a speed which is proportional to their distance from us. This recessional velocity is called the *Hubble flow* after its discoverer Edwin Hubble. This kind of expansion is easily visualized in the 'balloon' model (see Fig. 12.1). Paint regularly spaced dots on a spherical balloon and then inflate it. As it grows, the distance on the surface of the balloon between any two points grows at a rate proportional to that distance. Therefore *any* point will see all other points receding at a rate proportional to their distance. This proportionality preserves the homogeneity of the distribution of dots with time. It means that our location in the universe is not special, even though we appear to see everything else receding away from us. We are no more at the 'center' of the cosmological expansion than any other point is. The Hubble flow is compatible with the *Copernican Principle*, the idea that the universe does not revolve around (or expand away from) our particular location.

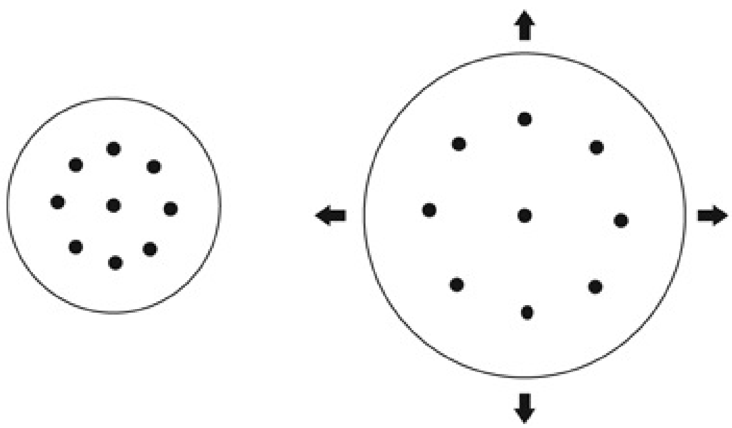


Figure 12.1 As the figure is magnified, all relative distances increase at a rate proportional to their magnitudes.

The Hubble expansion gives another opportunity for anisotropy. The universe would be homogeneous and anisotropic if every point saw a recessional velocity larger in, say, the x direction than in the y direction. In our model, this would happen if the balloon were an ellipsoid; to keep its shape it would have to expand faster along its longest axis than along the others. Our universe does not have any measurable velocity anisotropy. Because of this extraordinary simplicity, we can describe the relation between recessional velocity and distance with a single constant of proportionality H:

$$v = Hd \tag{12.1}$$

Astronomers call H *Hubble's parameter*. Its present value is called *Hubble's constant*, H_0. The value of H_0 is measured – by methods we discuss below – to be $H_0 = (71 \pm 4)$ km s^{-1}Mpc^{-1} in astronomers' peculiar but useful units. To get its value in normal units, convert 1 Mpc to 3.1×10^{22} m to get $H_0 = (2.3 \pm 0.1) \times 10^{-18}$ s^{-1}. In geometrized units, found by dividing by c, this is $H_0 = (7.7 \pm 0.4) \times 10^{-27}$ m^{-1}. Associated with the Hubble constant is the *Hubble time* $t_H = H_0^{-1} = (4.3 \pm 0.2) \times 10^{17}$ s. This is about 14 billion years, and is the time-scale for the cosmological expansion. The age of the universe will not exactly be this, since in the past the expansion speed varied, but this gives the order of magnitude of the time that has been available for the universe as we see it to have evolved.

We may object that the above discussion ignores the relativity of simultaneity. If the universe is changing in time – expanding – then it may be possible to find *some* definition of time such that hypersurfaces of constant time are homogeneous and isotropic, but this would not be true for other choices of a time coordinate. Moreover, Eq. (12.1) cannot be exact since, for $d > 1.3 \times 10^{26}$ m = 4200 Mpc, the velocity exceeds the velocity of light! This objection is right on both counts. Our discussion was a *local* one (applicable for recessional velocities $\ll 1$) and took the point of view of a particular observer, ourselves. Fortunately, the cosmological expansion is slow, so that over distances of 1000 Mpc, large enough to study the average properties of the homogeneous universe, the velocities are essentially nonrelativistic. Moreover, the average random velocities of galaxies relative to their near neighbors is typically less than 100 km s^{-1}, which is certainly nonrelativistic, and is much smaller than the systematic expansion speed over cosmological distances.

Therefore, the correct relativistic description of the expanding universe is that, in our neighborhood, there exists a *preferred choice of time,* whose hypersurfaces are homogeneous and isotropic, and with respect to which Eq. (12.1) is valid in the local inertial frame of *any* observer who is at rest with respect to these hypersurfaces at *any* location.

The existence of a preferred cosmological reference frame may at first seem startling: did we not introduce special relativity as a way to get away from special reference frames? There is no contradiction: the laws of physics themselves are invariant under a change of observer. But there is only one universe, and its physical make-up defines a convenient reference frame. Just as when studying the solar system it would be silly for us to place the origin of our coordinate system at, say, the position of Jupiter on 1 January 1900, so too would it be silly for us to develop the theory of cosmology in a frame that does not take advantage of the simplicity afforded by the large-scale homogeneity. From now on we will, therefore, work in the cosmological reference frame, with its preferred definition of time.

Models of the universe: the cosmological principle

If we are to make a large-scale model of the universe, we must make some assumption about regions that we have no way of seeing now because they are too distant for our telescope. We should in fact distinguish two different inaccessible regions of the universe.

The first inaccessible region is the region which is so distant that no information (traveling on a null geodesic) could reach us from it no matter how early this information began traveling. This region is everything that is outside our past light-cone. Such a region usually exists if the universe has a finite age, as ours does (see Fig. 12.2). This 'unknown' region is unimportant in one respect: what happens there has no effect on the interior of our past light cone, so how we incorporate it into our model universe has no effect on the way the model describes our observable history. On the other hand, our past light cone is a kind of horizon, which is called the *particle horizon:* as time passes, more and more of the previously unknown region enters the interior of our past light cone and becomes observable. So the unknown regions across the particle horizon can have a real influence on our future. In this sense, cosmology is a retrospective science: it reliably helps us understand only our past.

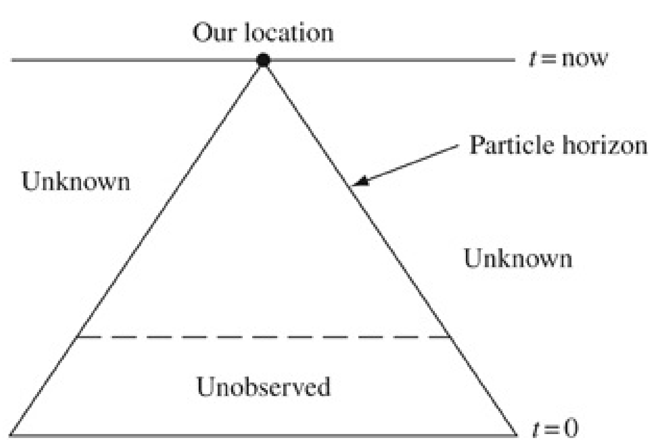


Figure 12.2 Schematic spacetime diagram showing the past history of the Universe, back to t = 0. The 'unknown' regions have not had time to send us

information; the ‘unobserved’ regions are obscured by intervening matter.

It must be acknowledged, however, that if information began coming in tomorrow that yesterday’s ‘unknown’ region was in fact very different from the observed universe, say highly inhomogeneous, then we would be posed difficult physical and philosophical questions regarding the apparently special nature of our history until this moment. It is to avoid these difficulties that we usually assume that the unknown regions are very like what we observe, and in particular are homogeneous and isotropic. Notice that there are very good reasons for adopting this idea. Consider, in Fig. 12.2, two hypothetical observers within our own past light cone, but at such an early time in the evolution of the universe that their own past light-cones are disjoint. Then they are outside each other’s particle horizon. But we can see that the physical conditions near each of them are very similar: we can confirm that if they apply the principle that regions outside their particle horizons are similar to regions inside, then they would be right! It seems unreasonable to expect that if this principle holds for other observers, then it will not also hold for us.

This modern version of the Copernican Principle is called the *Cosmological Principle*, or more informally the *Assumption of Mediocrity*, the ordinary-ness of our own location in the universe. It is, mathematically, an extremely powerful (i.e. restrictive) assumption. We shall adopt it, but we should bear in mind that predictions about the *future* depend strongly on the assumption of mediocrity.

The second inaccessible region is that part of the interior of our past light cone which our instruments cannot get information about. This includes galaxies so distant that they are too dim to be seen; processes that give off radiation – like gravitational waves – which we have not yet been able to detect; and events that are masked from view, such as those which emitted electromagnetic radiation before the epoch of decoupling (see below) when the universe ceased to be an ionized plasma and became transparent to electromagnetic waves. The limit of decoupling is sometimes called our *optical horizon* since no light reaches us from beyond it (from earlier times). But gravitational waves do propagate freely before this, so eventually we will begin to make observations across this ‘horizon’: the optical horizon is not a fundamental limit in the way the particle horizon is.

Cosmological metrics

The metric tensor that represents a cosmological model must incorporate the observed homogeneity and isotropy. We shall therefore adopt the following idealizations about the universe: (i) spacetime can be sliced into hypersurfaces of constant time which are perfectly homogeneous and isotropic; and (ii) the mean rest frame of the galaxies agrees with this definition of simultaneity. Let us next try to simplify the problem as much as possible by adopting *comoving coordinates:* each galaxy is idealized as having no random velocity, and we give each galaxy a fixed set of coordinates $\{x^i, i = 1, 2, 3\}$. We choose our time coordinate t to be proper time for each galaxy. The expansion of the universe – the change of proper distance between galaxies – is represented by time-dependent metric coefficients. Thus, if at one moment, t_0, the hypersurface of constant time has the line element

$$dl^2(t_0) = h_{ij}(t_0)\, dx^i\, dx^j \tag{12.2}$$

(these hs have nothing to do with linearized theory), then the expansion of the hypersurface can be represented by

$$\begin{aligned} dl^2(t_1) &= f(t_1, t_0) h_{ij}(t_0)\, dx^i\, dx^j \\ &= h_{ij}(t_1)\, dx^i\, dx^j. \end{aligned} \tag{12.3}$$

This form guarantees that all the h_{ij}s increase at the same rate; otherwise the expansion would be anisotropic (see Exer. 4, § 12.6). In general, then, Eq. (12.2) can be written

$$dl^2(t) = R^2(t) h_{ij}\, dx^i\, dx^j, \tag{12.4}$$

where R is an overall scale factor which equals one at t_0, and where h_{ij} is a constant metric equal to that of the hypersurface at t_0. We shall explore what form h_{ij} can take in detail in a moment.

First we extend the constant-time hypersurface line element to a line element for the full spacetime. In general, it would be

$$ds^2 = -dt^2 + g_{0i}\, dt\, dx^i + R^2(t) h_{ij}\, dx^i\, dx^j, \tag{12.5}$$

where $g_{00} = -1$, because t is proper time along a line $\mathrm{d}x^i = 0$. However, if the definition of simultaneity given by t = const. is to agree with that given by the local Lorentz frame attached to a galaxy (idealization (ii) above), then $\vec{e}_0$ must be orthogonal to $\vec{e}_i$ in our comoving coordinates. This means that $g_{0i} = \vec{e}_0 \cdot \vec{e}_i$ must vanish, and we get

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + R^2(t)h_{ij}\,\mathrm{d}x^i\,\mathrm{d}x^j. \tag{12.6}$$

What form can h_{ij} take? Since it is isotropic, it must be spherically symmetric about the origin of the coordinates, which can of course be chosen to be located at any point we like. When we discussed spherical stars we showed that a spherically symmetric metric always has the line element (last part of Eq. (10.5))

$$\mathrm{d}l^2 = \mathrm{e}^{2\Lambda(r)}\,\mathrm{d}r^2 + r^2\mathrm{d}\Omega^2. \tag{12.7}$$

This form of the metric implies only isotropy about one point. We want a stronger condition, namely that the metric is homogeneous. A necessary condition for this is certainly that the Ricci scalar curvature of the three-dimensional metric, R_{ii}, must have the same value at every point: every scalar must be independent of position at a fixed time. We will show below, remarkably, that this is sufficient as well, but for now we just treat it as the next constraint we place on the metric in Eq. (12.7). We can calculate $R^i{}_i$ using Exer. 35 of § 6.9. Alternatively, we can use Eqs. (10.15)–(10.17) of our discussion of spherically symmetric spacetimes in Ch. 10, realizing that G_{ij} for the line element, Eq. (12.7), above is obtainable from G_{ij} for the line element, Eq. (10.7), of a spherical star by setting Φ to zero. We get

$$\begin{aligned} G_{rr} &= -\frac{1}{r^2}\mathrm{e}^{2\Lambda}(1 - e^{-2\Lambda}), \\ G_{\theta\theta} &= -r\,\mathrm{e}^{-2\Lambda}\Lambda', \\ G_{\phi\phi} &= \sin^2\theta\, G_{\theta\theta}. \end{aligned} \tag{12.8}$$

The trace of this tensor is also a scalar, and must also therefore be constant. So instead of computing the Ricci scalar curvature, we simply require that the trace G of the three-dimensional Einstein tensor be a constant. (In fact, this trace is just −1/2 of the Ricci scalar.) The trace is

$$
\begin{aligned}
G &= G_{ij}g^{ij} \\
&= -\frac{1}{r^2}\,e^{2\Lambda}(1-e^{-2\Lambda})\,e^{-2\Lambda} - 2r\,e^{-2\Lambda}\Lambda' r^{-2} \\
&= -\frac{1}{r^2} + \frac{1}{r^2}e^{-2\Lambda}(1-2r\Lambda') \\
&= -\frac{1}{r^2}[1-(r\,e^{-2\Lambda})'].
\end{aligned}
\tag{12.9}
$$

Demanding homogeneity means setting G to some constant κ:

$$
\kappa = -\frac{1}{r^2}[r(1-e^{-2\Lambda})]'. \tag{12.10}
$$

This is easily integrated to give

$$
g_{rr} = e^{2\Lambda} = \frac{1}{1+\frac{1}{3}\kappa r^2 - A/r}, \tag{12.11}
$$

where A is a constant of integration. As in the case of spherical stars, we must demand local flatness at $r = 0$ (compare with § 10.5): $g_{rr}(r = 0) = 1$. This implies $A = 0$. Defining the more conventional *curvature constant* $k = -\kappa/3$ gives

$$
\begin{aligned}
g_{rr} &= \frac{1}{1-kr^2} \\
dl^2 &= \frac{dr^2}{1-kr^2} + r^2\,d\Omega^2.
\end{aligned}
\tag{12.12}
$$

We have not yet proved that this space is isotropic about every point; all we have shown is that Eq. (12.12) is the unique space which satisfies the necessary condition that this curvature scalar be homogeneous. Thus, if a space that is isotropic and homogeneous exists at all, it must have the metric, Eq. (12.12), for at least some k.

In fact, the converse *is* true: the metric of Eq. (12.12) is homogeneous and isotropic for *any* value of k. We will demonstrate this explicitly for positive, negative, and zero k separately in the next paragraph. General proofs not depending on the sign of k can be found in, for example, Weinberg (1972) or Schutz (1980*b*). Assuming this result for the time being, therefore, we conclude that the full cosmological spacetime has the metric

$$ds^2 = -dt^2 + R^2(t)\left[\frac{dr^2}{1-kr^2} + r^2\, d\Omega^2\right]. \tag{12.13}$$

This is called the *Robertson–Walker metric*. Notice that we can, without loss of generality, scale the coordinate r in such a way as to make k take one of the three values +1, 0, −1. To see this, consider for definiteness $k = -3$. Then redefine $\tilde{r} = \sqrt{3}r$ and $\tilde{R} = 1/\sqrt{3}R$, and the line element becomes

$$ds^2 = -dt^2 + \tilde{R}^2(t)\left[\frac{d\tilde{r}^2}{1-\tilde{r}^2} + \tilde{r}^2\, d\Omega^2\right]. \tag{12.14}$$

What we cannot do with this rescaling is change the sign of k. Therefore there are only three spatial hypersurfaces we need consider: $k = (-1, 0, 1)$.

Three types of universe

Here we prove that all three kinds of hypersurfaces represent homogenous and isotropic metrics that have different large-scale geometries. Consider first $k = 0$. Then, at any moment t_0, the line element of the hypersurface (setting $dt = 0$) is

$$dl^2 = R^2(t_0)\left[dr^2 + r^2\, d\Omega^2\right] = d(r')^2 + (r')^2\, d\Omega, \tag{12.15}$$

with $r' = R(t_0)r$. (Remember that $R(t_0)$ is constant on the hypersurface.) This is clearly the metric of flat Euclidean space. This is the *flat* Robertson–Walker universe. That it is homogeneous and isotropic is obvious.

Consider, next, $k = +1$. Let us define a new coordinate $\chi(r)$ such that

$$d\chi^2 = \frac{dr^2}{1 - r^2} \tag{12.16}$$

and $\chi = 0$ where $r = 0$. This integrates to

$$r = \sin\chi, \tag{12.17}$$

so that the line element for the space $t = t_0$ is

$$dl^2 = R^2(t_0)[d\chi^2 + \sin^2\chi\,(d\theta^2 + \sin^2\theta\, d\phi^2)]. \tag{12.18}$$

We showed in Exer. 33, § 6.9, that this is the metric of a three-sphere of radius $R(t_0)$, *i.e.* of the set of points in four-dimensional Euclidean space that are all at a distance $R(t_0)$ from the origin. This model is called the *closed*, or *spherical* Robertson–Walker metric and the balloon analogy of cosmological expansion (Fig. 12.1) is particularly appropriate for it. It is clearly homogeneous and isotropic: no matter where we stand on the three-sphere, it looks the same in all directions. Remember that the fourth spatial dimension – the radial direction to the center of the three-sphere – has *no* physical meaning to us: all our measurements are confined to our three-space so we can have no physical knowledge about the properties or even the existence of that dimension. At this point we should perhaps think of it as simply a tool for making it easy to visualize the three-sphere, not as an extra real dimension, although see the final paragraph of this chapter for a potentially different point of view on this.

The final possibility is $k = -1$. An analogous coordinate transformation (Exer. 8, § 12.6) gives the line element

$$\mathrm{d}l^2 = R^2(t_0)(\mathrm{d}\chi^2 + \sinh^2 \chi \, \mathrm{d}\Omega^2). \tag{12.19}$$

This is called the *hyperbolic*, or *open*, Robertson–Walker model. Notice one peculiar property. As the proper radial coordinate χ increases away from the origin, the circumferences of spheres increase as $\sinh \chi$. Since $\sinh \chi > \chi$ for all $\chi > 0$, it follows that these circumferences increase *more* rapidly with proper radius than in flat space. For this reason this hypersurface is *not* realizable as a three-dimensional hypersurface in a fouror higher-dimensional Euclidean space. That is, there is no picture which we can easily draw such as that for the three-sphere. The space is call 'open' because, unlike for $k = +1$, circumferences of spheres increase monotonically with χ: there is no natural end to the space.

In fact, as we show in Exer. 8, § 12.6, this geometry is the geometry of a hypersurface embedded in *Minkowski* spacetime. Specifically, it is a hypersurface of events that all have the same timelike interval from the origin. Since this hypersurface has the same interval from the origin in any Lorentz frame (intervals are Lorentz invariant), this hypersurface is indeed homogeneous and isotropic.

Cosmological redshift as a distance measure

When studying small regions of the universe around the Sun, astronomers measure proper distances to stars and other objects and express them in parsecs, as we have seen, or in the multiples kpc and Mpc. But if the object is at a cosmological distance in a universe that is expanding, then what we mean by distance is a little ambiguous, due to the long time it takes light to travel from the object to us. Its separation from our location when it emitted the light that we receive today may have been much less than its separation at present, *i.e.* on the present hypersurface of constant time. Indeed, the object may not even exist any more: all we know about it is that it existed at the event on our past light-cone when it emitted the light we receive today. But between then and now it might have exploded, collapsed, or otherwise changed dramatically. So the notion of the separation between us and the object *now* is not as important as it might be for local measurements.

Instead, astronomers commonly use a different measurement of separation: the redshift z of the spectrum of the light emitted by the object, let us say a galaxy. In an expanding universe that follows Hubble's Law, Eq. (12.1), the further away the galaxy is, the faster it is receding from us, so the redshift is a nice monotonic measure of separation: larger redshifts imply larger distances. Of course, as we noted in the discussion following Eq. (12.1), the galaxy's redshift contains an element due to its random local velocity; over cosmological distances this is a small uncertainty, but for the nearby parts of the universe astronomers use conventional distance measures, mainly Mpc, instead of redshift.

To compute the redshift in our cosmological models, let us assume that the galaxy has a fixed coordinate position on some hypersurface at the cosmological time t at which it emits the light we eventually receive at time t_0. Recall our discussion of conserved quantities in § 7.4: if the metric is independent of a coordinate, then the associated covariant component of momentum is constant along a geodesic. In the cosmological case, the homogeneity of the hypersurfaces ensures that the covariant components of the spatial momentum of the photon emitted by our galaxy are constant along its trajectory. Suppose that we place ourselves at the origin of the cosmological coordinate system (since the cosmology is homogeneous, we can put the origin anywhere we like), so that light travels along a radial line θ = const., ϕ = const. to us. In each of the cosmologies the line-element restricted to the trajectory has the form

$$0 = -\mathrm{d}t^2 + R^2(t)\mathrm{d}\chi^2. \tag{12.20}$$

(To get this for the flat hypersurfaces, simply rename the coordinate r in the first part of Eq. (12.15) to χ.) It follows that the relevant conserved quantity for the photon is p_χ . Now, the cosmological time coordinate t is proper time, so the energy as measured by a local observer at rest in the cosmology anywhere along the trajectory is $-p^0$. We argue in Exer. 9, § 12.6, that conservation of p_χ implies that p^0 is inversely proportional to $R(t)$. It follows that the wavelength as measured locally (in proper distance units) is proportional to $R(t)$, and hence that the redshift z of a photon emitted at time t and observed by us at time t_0 is given by

$$1 + z = R(t_0)/R(t). \tag{12.21}$$

It is important to keep in mind that this is just the cosmological part of any overall redshift: if the source or observer is moving relative to the cosmological rest frame, then there will be a further factor of $1 + z_{\text{motion}}$ multiplied into the right-hand-side of Eq. (12.24).

We now show that the *Hubble parameter* $H(t)$ is the instantaneous relative rate of expansion of the universe at time t:

$$H(t) = \frac{\dot{R}(t)}{R(t)}. \tag{12.22}$$

Our galaxy at a fixed coordinate location χ is carried away from us by the cosmological expansion. At the present time t_0 its proper distance d from us (in the constant-time hypersurface) is the same for each of the cosmologies when expressed in terms of χ:

$$d_0 = R(t_0)\chi. \tag{12.23}$$

It follows by differentiating this that the current rate of change of proper distance between the observer at the origin and the galaxy at fixed χ is

$$v = (\dot{R}/R)_0 d_0 = H_0 d_0, \tag{12.24}$$

where H_0 is the present value of the Hubble parameter. By comparison with Eq. (12.1), we see that this is just the present value of the Hubble parameter $\dot{R}/R$. We show in Exer. 10, § 12.6 that this velocity is just $v = z$, which is what is required to give the redshift z, provided the galaxy is not far away. In our cosmological neighborhood, therefore, the cosmological redshift is a true Doppler shift. Moreover, the redshift is proportional to proper distance in our neighborhood, with the Hubble constant as the constant of proportionality. We shall now investigate how various measures of distance depend on redshift when we leave our cosmological neighborhood.

The scale factor of the Universe $R(t)$ is related to the Hubble parameter by Eq. (12.22). Integrating this for R gives

$$R(t) = R_0 \exp\left[\int_{t_0}^{t} H(t')\mathrm{d}t'\right]. \tag{12.25}$$

The Taylor expansion of this is

$$R(t) = R_0[1 + H_0(t - t_0) + \tfrac{1}{2}(H_0^2 + \dot{H}_0)(t - t_0)^2 + \cdots], \tag{12.26}$$

where subscripted zeros denote quantities evaluated at t_0. The time-derivative of the Hubble parameter contains information about the acceleration or deceleration of the expansion. Cosmologists sometimes replace $\dot{H}_0$ with the dimensionless *deceleration parameter*, defined as

$$q_0 = -R_0\ddot{R}_0/\dot{R}_0^2 = -\left(1 + \dot{H}_0/H_0^2\right). \tag{12.27}$$

The minus sign in the definition and the name 'deceleration parameter' reflect the assumption, when this parameter was first introduced, that gravity would be slowing down the cosmological expansion, so that q_0 would be positive. However, astronomers now believe that the universe is accelerating, so the idea of a 'deceleration parameter' has gone out of fashion. Nevertheless, any formula containing $\dot{H}_0$ can be converted to one in terms of q_0 and vice-versa.

What does Hubble's law, Eq. (12.1), look like to this accuracy? The recessional velocity v is deduced from the redshift of spectral lines, so it is more convenient to work directly with the redshift. Combining Eq. (12.25) with Eq. (12.21) we get

$$1 + z(t) = \exp\left[-\int_{t_0}^{t} H(t')\mathrm{d}t'\right]. \tag{12.28}$$

The Taylor expansion of this is

$$z(t) = H_0(t_0 - t) + \tfrac{1}{2}(H_0^2 - \dot{H}_0)(t_0 - t)^2 + \cdots. \tag{12.29}$$

This is not directly useful yet, since we have no independent information about the time t at which a galaxy emitted its light. Perhaps Eq. (12.29) is more useful when inverted to give an expansion for the look-back proper time to an event with redshift z:

$$t_0 - t(z) = H_0^{-1}\left[z - \tfrac{1}{2}(1 - \dot{H}_0/H_0^2)z^2 + \cdots\right]. \tag{12.30}$$

From the simple expansion

$$H(t) = H_0 + \dot{H}_0(t - t_0) + \cdots ,$$

we can substitute in the first term of the previous equation and get an expansion for H as a function of z:

$$H(z) = H_0 \left(1 - \frac{\dot{H}_0}{H_0^2} z + \cdots \right). \tag{12.31}$$

Note that Eq. (12.28) can also be inverted to give the exact and very simple relation

$$H(t) = -\frac{\dot{z}}{1 + z}. \tag{12.32}$$

Although cosmology is self-consistent only within a relativistic framework, it is nevertheless useful to ask how the expansion of the universe looks in Newtonian language. We imagine a spherical region uniformly filled with galaxies, starting at some time with radially outward velocities that are proportional to the distance from the center of the sphere. If we are not near the edge – and of course the edge may be much too far away for us to see today – then we can show that the expansion is homogeneous and isotropic about every point. The galaxies just fly away from one another, and the Hubble constant is the scale for the initial velocity: it is the radial velocity per unit distance away from the origin. The problem with this Newtonian model is not that it cannot describe the local state of the universe, it is that, with gravitational forces that propagate instantaneously, the dynamics of any bit of the universe depends on the structure of this cloud of galaxies arbitrarily far away. Only in a relativistic theory of gravity can we make sense of the dynamical evolution of the universe. This is a subject we will study below.

When light is redshifted, it loses energy. Where does this energy go? The fully relativistic answer is that it just goes away: since the metric depends on time, there is no conservation law for energy along a geodesic. Interestingly, in the Newtonian picture of the universe just described, the redshift is just caused by the different velocities of the diverging galaxies relative to one another. As the photon moves outward in the expanding cloud, it finds itself passing galaxies that are moving faster and faster relative to the center. It is not surprising that they measure the energy of the photon to be smaller and smaller as it moves outwards.

Cosmography: measures of distance in the universe

Cosmography refers to the description of the expansion of the universe and its history. In cosmography we do not yet apply the Einstein equations to explain the motion of the universe, instead we simply measure its expansion history. The language of cosmography is the language of distance measures and the evolution of the Hubble parameter.

By analogy with Eq. (12.1), we would like to replace t in Eq. (12.29) with distance. But what measure of distance is suitable over vast cosmological separations? Not coordinate distance, which would be unmeasurable. What about proper distance? The proper distance between the events of emission and reception of the light is zero, since light travels on null lines. The proper distance between the emitting galaxy and us at the present time is also unmeasurable: in principle, the galaxy may not even exist now, perhaps because of a collision with another galaxy. To get out of this difficulty, let us ask how distance crept into Eq. (12.1) in the first place.

Distances to nearby galaxies are almost always inferred from luminosity measurements. Consider an object whose distance d is known, which is at rest, and which is near enough to us that we can assume that space is Euclidean. Then a measurement of its flux F leads to an inference of its absolute luminosity:

$$L = 4\pi d^2 F \tag{12.33}$$

Alternatively, if L is known, then a measurement of F leads to the distance d. The role of d in Eq. (12.1) is, then, as a replacement for the observable $(L/F)^{1/2}$.

Astronomers have used brightness measurements to build up a carefully calibrated *cosmological distance ladder* to measure the scale of the universe. For each step on this ladder they identify what is called a *standard candle*, which is a class of objects whose absolute luminosity L is known (say from a theory of their nature or from reliably calibrated distances to nearby examples of this object). As their ability to see to greater and greater distances has developed, astronomers have found new standard candles that they could calibrate from previous ones but that were bright enough to be seen to greater distances than the previous ones. The distance ladder starts at the nearest stars, the distances to which can be measured by parallax (independently of luminosity), and continues all the way to very distant high-redshift galaxies.

In the spirit of such measurements, cosmologists define the *luminosity distance* d_L to any object, no matter how distant, by inverting Eq. (12.33):

$$d_{\mathrm{L}} = \left(\frac{L}{4\pi F}\right)^{1/2}. \tag{12.34}$$

The luminosity distance is often the observable that can be directly measured by astronomers: if the intrinsic luminosity L is known or can be inferred, then a measurement of its brightness F determines the luminosity distance. The luminosity distance is the proper distance the object would have in a Euclidean universe if it were at rest with respect to us, if it had an intrinsic luminosity L, and if we received an energy flux F from it. However, in an expanding cosmology this will not generally be the proper distance to the object.

We shall now find the relation between luminosity distance and the cosmological scales we have just introduced. Consider an object emitting with luminosity L at a time t_e. What flux do we receive from it at the later time t_0? Suppose for simplicity that the object gives off only photons of frequency ν_e at time t_e. (This frequency will drop out in the end, so our result will be perfectly general.) In a small interval of time δt_e the object emits

$$N = L\delta t_{\mathrm{e}}/h\nu_{\mathrm{e}} \tag{12.35}$$

photons in a spherically symmetric manner. To find the flux we receive, we must calculate the area of the sphere that these photons occupy at the time we observe them.

We place the object at the origin of the coordinate system, and suppose that we sit at coordinate position r in this system, as given in Eq. (12.13). Then when the photons reach our coordinate distance from the emitting object, the proper area of the sphere they occupy is given by integrating over the sphere the solid-angle part of the line-element in Eq. (12.13), which is $R_0^2 r^2 \mathrm{d}\Omega^2$. The integration is just on the spherical angles and produces the area:

$$A = 4\pi R_0^2 r^2. \tag{12.36}$$

Now, the photons have been redshifted by the amount $(1 + z) = R_0/R(t_e)$ to frequency ν_0:

$$h\nu_0 = h\nu_e/(1+z). \tag{12.37}$$

Moreover, they arrive spread out over a time δt_0, which is also stretched by the redshift:

$$\delta t_0 = \delta t_e(1+z). \tag{12.38}$$

The energy flux at the observation time t_0 is thus $Nh\nu_0/(A\delta t_0)$, from which it follows that

$$F = L/A(1+z)^2. \tag{12.39}$$

From Eq. (12.34), we then find

$$d_L = R_0 r(1+z). \tag{12.40}$$

To use this, we need to know the comoving source coordinate location r as a function of the redshift z of the photon the source emitted. This comes from solving the equation of motion of the photon. In this case, all we have to do is use Eq. (12.13) with $ds^2 = 0$ (a photon world line) and $d\Omega^2 = 0$ (photon traveling on a radial line from its emitter to the observer at the center of the coordinates). This leads to the differential equation

$$\frac{dr}{\left(1-kr^2\right)^{1/2}} = -\frac{dt}{R(t)} = \frac{dz}{R_0 H(z)}, \tag{12.41}$$

where the last step follows from differentiating Eq. (12.21). This equation involves the curvature parameter k, but for small r and z the curvature will come into the solution only at second order. If we ignore this at present and work only to first order beyond the Euclidean relations, it is not hard to show that

$$d_L = R_0 r(1+z) = \left(\frac{z}{H_0}\right)\left[1 + \left(1 + \frac{1}{2}\frac{\dot{H}_0}{H_0^2}\right)z\right] + \cdots . \tag{12.42}$$

If we can measure the luminosity distances and redshifts of a number of objects, then we can in principle measure $\dot{H}_0$. Measurements of this kind led to the discovery of the accelerating expansion of the universe (below).

Another convenient measure of distance is the *angular diameter distance*. This is based on another way of measuring distances in a Euclidean space: the angular size θ of an object at a distance d can be inferred if we know the proper diameter D of the object transverse to the line of sight, $\theta = D/d$. This leads to the definition of the angular diameter distance d_A to an object anywhere in the universe:

$$d_A = D/\theta. \tag{12.43}$$

The dependence of d_A on redshift z is explored in Exer. 12, § 12.6. The result is

$$d_A = R_e r = (1+z)^{-2} d_L, \tag{12.44}$$

where R_e is the scale factor of the universe when the photon was emitted. The analogous expression to Eq. (12.42) is

$$d_A = R_0 r/(1+z) = \left(\frac{z}{H_0}\right)\left[1 + \left(-1 + \frac{1}{2}\frac{\dot{H}_0}{H_0^2}\right)z\right] + \cdots . \tag{12.45}$$

There are situations where we have in fact an estimate of the comoving diameter D of an emitter. In particular, the temperature irregularities in maps of the cosmic microwave background radiation (see below) have a length scale that is determined by the physics of the early universe.

Although we have provided small-z expansions for many interesting measures, it is important to bear in mind that astronomers today can observe objects out to very high redshifts. Some galaxies and quasars are known at redshifts greater than $z = 6$. The cosmological microwave background, which we will discuss below, originated at redshift $z \sim 1000$, and is our best tool for understanding the Big Bang. Even so, the universe was already some 300 000 years old at that redshift. Sometime in the future, gravitational wave detectors may detect random radiation from the Big Bang itself, originating when the universe was only a fraction of a second old.

The derivation of Eq. (12.42) illustrates a point which we have encountered before: in the attempt to translate the nonrelativistic formula $v = Hd$ into relativistic language, we were forced to re-think the meaning of all the terms in the equations and to go back to the quantities we can directly measure. If the study of GR teaches us only one thing, it should be that physics rests ultimately on *measurements*: concepts like distance, time, velocity, energy, and mass are derived from measurements, but they are often not the quantities directly measured, and our assumptions about their global properties must be guided by a careful understanding of how they are related to measurements.

The universe is accelerating!

The most remarkable cosmographic result since Hubble's original work was the discovery that the expansion of the universe is not slowing down, but rather speeding up. This was done by essentially making a plot of the luminosity distance against redshift, but where luminosities are given in magnitudes. This is called the *magnitude–redshift* diagram, and we derive its low-z expansion in Exer. 13, § 12.6. Two teams of astronomers, called the High-Z Supernova Search Team (Riess *et al.*, 1998) and the Supernova Cosmology Project (Perlmutter *et al.*, 1999), respectively, used supernova explosions of Type Ia as standard candles out to redshifts of order 1. Although there was considerable scatter among the data points, both teams found that the best fit to the data was a universe that was speeding up and not slowing down. The data from the High-Z Team are shown in Fig. 12.3. See Filippenko (2008) for a full discussion.

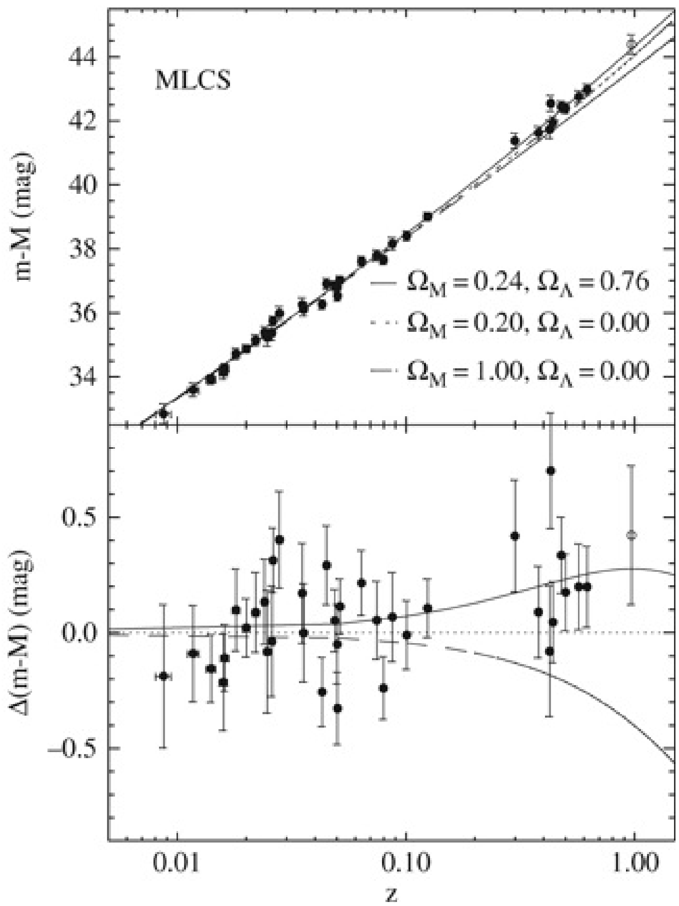


Figure 12.3 The trend of luminosity versus redshift for Type Ia supernovae is fit best with an accelerating universe. The lower part of this curve determines H_0, the upper part demonstrates acceleration. (High-Z Supernova Search Team: Riess, *et al*, 1998.)

The top diagram shows the flux (magnitude) measurement for each of the supernovae in the sample, along with error bars. The trend seems to curve upwards, meaning that at high redshifts the supernovae are dimmer than expected. This would happen if the universe were speeding up, because the supernovae would simply be further away than expected. Three possible fits are shown, and the best one has a large positive cosmological constant, which we shall see below is the simplest way, within Einstein's equations, that we can accommodate acceleration. The lower diagram shows the same data but plotting only the residuals from the fit to a flat universe. This shows more clearly how the data favor the curve for the accelerating universe.

These studies were the first strong evidence for acceleration, but by now there are several lines of investigation that lead to the same conclusion. Astronomers initially resisted the conclusion, because it undermines a basic assumption we have always made about gravity, that it is universally attractive. If the energy density of the universe exerts attractive gravity, the expansion should be slowing down. Instead it speeds up. What can be the cause of this repulsion? We shall return to this question repeatedly through the rest of this chapter.

12.3 Cosmological dynamics: understanding the expanding universe

In the last section we saw how to describe a homogeneous and isotropic universe and how to measure its expansion. In order to study the evolution of the universe, to understand the creation of the huge variety of structures that we see, and indeed to make sense of the accelerating expansion measured today, we have to apply Einstein's equations to the problem, and to marry them with enough physics to explain what we see. In this section we will study Einstein's equations, with relatively simple perfect-fluid physics and with the cosmological constant that seems to be implied by the expansion. In the following sections we will study more and more of cosmological physics.

Dynamics of Robertson–Walker universes: Big Bang and dark energy

We have seen that a homogeneous and isotropic universe must be described by one of the three Robertson–Walker metrics given by Eq. (12.13). For each choice of the curvature parameter $k = (-1, 0, +1)$, the evolution of the universe depends on just one function of time, the scale factor $R(t)$. Einstein's equations will determine its behavior.

As in earlier chapters, we idealize the universe as filled with a homogeneous perfect fluid. The fluid must be at rest in the preferred cosmological frame, for otherwise its velocity would allow us to distinguish one spatial direction from another: the universe would not be isotropic. Therefore, the stress-energy tensor will take the form of Eq. (4.36) in the cosmological rest frame. Because of homogeneity, all fluid properties depend only on time: $\rho = \rho(t)$, $p = p(t)$, *etc.*

First we consider the equation of motion for matter, $T^{\mu\nu}{}_{;\nu} = 0$, which follows from the Bianchi identities of Einstein's field equations. Because of isotropy, the spatial components of this equation must vanish identically. Only the time component $\mu = 0$ is nontrivial. It is easy to show (see Exer. 14, § 12.6) that it gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\rho R^3\right) = -p\frac{\mathrm{d}}{\mathrm{d}t}\left(R^3\right), \tag{12.46}$$

where $R(t)$ is the cosmological expansion factor. This is easily interpreted: R^3 is proportional to the volume of any fluid element, so the left-hand side is the rate of change of its total energy, while the right-hand side is the work it does as it expands ($-p\ \mathrm{d}\ v$).

There are two simple cases of interest, a *matter-dominated cosmology* and a *radiation-dominated cosmology*. In a matter-dominated era, which includes the present epoch, the main energy density of the cosmological fluid is in cold nonrelativistic matter particles, which have random velocities that are small and which therefore behave like dust: $p = 0$. So we have

$$\text{Matter-dom}: \quad \frac{\mathrm{d}}{\mathrm{d}t}\left(\rho R^3\right) = 0. \tag{12.47}$$

In a radiation-dominated era (as we shall see, in the early universe), the principal energy density of the cosmological fluid is in radiation or hot, highly relativistic particles, which have an equation of state $p = \frac{1}{3}\rho$ (Exer. 22, § 4.10). Then we get

$$\text{Radiation-dom}: \quad \frac{\mathrm{d}}{\mathrm{d}t}(\rho R^3) = -\tfrac{1}{3}\rho\frac{\mathrm{d}}{\mathrm{d}t}(R^3), \tag{12.48}$$

or

$$\text{Radiation-dom}: \quad \frac{\mathrm{d}}{\mathrm{d}t}\left(\rho R^4\right) = 0. \tag{12.49}$$

The Einstein equations are also not hard to write down for this case. Isotropy will guarantee that $G_{tj} = 0$ for all j, and also that $G_{jk} \propto g_{jk}$. That means that only two components are independent, G_{tt} and (say) G_{rr}. But the Bianchi identity will provide a relationship between them, which we have already used in deriving the matter equation in the previous paragraph. (The same happened for the spherical star.) Therefore we only need compute one component of the Einstein tensor (see Exer. 16, § 12.6):

$$G_{tt} = 3(\dot{R}/R)^2 + 3k/R^2. \tag{12.50}$$

Therefore, besides Eqs. (12.47) or (12.49), we have only one further equation, the Einstein equation with cosmological constant Λ

$$G_{tt} + \Lambda g_{tt} = 8\pi T_{tt}. \tag{12.51}$$

Physicists today hope that they will eventually be able to compute the value of the cosmological constant from first principles in a consistent theory where gravity is quantized along with all the other fundamental interactions. From this point of view, the cosmological constant will represent just another contribution to the whole stress-energy tensor, which can be given the notation

$$T_{\Lambda}^{\alpha\beta} = -(\Lambda/8\pi)g^{\alpha\beta}. \tag{12.52}$$

From this point of view, the energy density and pressure of the cosmological constant 'fluid' are

$$\rho_\Lambda = \Lambda/8\pi, \qquad p_\Lambda = -\rho_\Lambda. \tag{12.53}$$

Cosmologists call ρ_Λ the *dark energy:* an energy that is not associated with any known matter field. Its associated dark pressure p_Λ has the opposite sign.

Physicists generally expect that the dark energy will be positive, so that most discussions of cosmology today are in the framework of $\Lambda \geq 0$. We will return below to the implications of the associated negative value for the dark pressure. Notice that, as the universe expands, the dark energy density and dark pressure remain *constant*. In these terms the *tt*-component of the Einstein equations can be written

$$\frac{1}{2}\dot{R}^2 = -\frac{1}{2}k + \frac{4}{3}\pi R^2(\rho_m + \rho_\Lambda), \tag{12.54}$$

where now we write ρ_m for the energy density of the matter (including radiation), to distinguish it from the dark energy density. This equation makes it easy to understand the observed acceleration of our universe. It appears (see below) that $k = 0$, or at least that the k-term is negligible. Then, since we are in a matter-dominated epoch, the term $R^2\rho_m$ decreases as R increases, while the term $R^2\rho_\Lambda$ increases rather strongly. Since today, as we shall see below, $\rho_\Lambda > \rho_m$, the result is that $\dot{R}$ increases as R increases. This trend must continue now forever, provided the acceleration is truly propelled by a cosmological constant, and not by some physical field that will go away later.

How, physically, can a positive dark energy density drive the universe into accelerated expansion? Is not positive energy gravitationally attractive, so would it not act to slow down the expansion? To answer this it is helpful to look at the spatial part of Einstein's equations, where there acceleration $\ddot{R}$ explicitly appears. Rather than derive this from the Christoffel symbols, we can use the fact that (as remarked above) it follows from the two basic equations we have already written down: Eq. (12.46) and the time-derivative of Eq. (12.54). In Exer. 17, § 12.6 we show that the combination of these two equations implies the following simple 'equation of motion' for the scale factor:

$$\frac{\ddot{R}}{R} = -\frac{4\pi}{3}(\rho + 3p), \tag{12.55}$$

where ρ and p are the *total* energy density and pressure, including both the normal matter and the dark energy.

The acceleration is produced, not by the energy density alone, but by $\rho + 3p$. We have met this combination before, in Exer. 20, § 8.6, where we called it the *active gravitational mass*. We showed there that, in general relativity, when pressure cannot be ignored, the source of the far-away Newtonian field is $\rho + 3p$, not just ρ. In the cosmological context, the same combination generates the cosmic acceleration. It is clear that the negative pressure associated with the cosmological constant can, if it is large enough, make this sum negative, and that is what drives the universe faster and faster. Einstein's gravity with a cosmological constant has a kind of in-built anti-gravity!

Notice that a negative pressure is not by any means unphysical. Negative stress is called *tension*, and in a stretched rubber band, for example, the component of the stress tensor along the band is negative. Interestingly, our analogy using a balloon to represent the expanding universe also introduces a negative pressure, the tension in the stretched rubber. What is remarkable about the dark energy is that its tension is so large, and it is isotropic. See Exer. 18, § 12.6 for a further discussion of the tension in this 'fluid'.

Eq. (12.54) is written in a form that suggests studying the expansion of the universe in a way analogous to the energy methods physicists use for particle motion (as we did for orbits in Schwarzschild in Ch. 11). The left-hand side looks like a 'kinetic energy' and the right-hand side contains a constant $(-k/2)$ that plays the role of the 'total energy' and a potential term proportional to $R^2(\rho_m + \rho_\Lambda)$, which depends on R explicitly and through ρ_m. The dynamics of R will be constrained by this energy equation.

We can use this constraint to explore what might happen to our universe in the far distant future, assuming of course the Cosmological Principle, that nothing significantly new comes over our particle horizon. If $\rho_\Lambda \geq 0$ (see above) and if the matter content of the universe also has positive energy density, then one conclusion from Eq. (12.54) is immediate: *an expanding hyperbolic universe (k = −1) will never stop expanding*. For the flat universe ($k = 0$), an expanding universe will also never stop if $\rho_\Lambda > 0$; however, if $\rho_\Lambda = 0$, then it could asymptotically slow down to a zero expansion rate as R approaches infinity, since the matter density will decrease at least as fast as R^{-3}. An expanding closed universe ($k = 1$) will, if $\rho_\Lambda = 0$, always reach a maximum expansion radius and then turn around and re-collapse, again because $R^2{}_{\rho m}$ decreases with R. A re-collapsing universe eventually reaches another singularity, called the *Big Crunch!* But if $\rho_\Lambda > 0$, then the ultimate fate of an expanding closed universe depends on the balance of ρ_m and ρ_Λ.

We can ask similar questions about the *history* of our universe: was there a Big

Bang, where the scale factor R had the value zero at a finite time in the past? First we consider for simplicity $\rho_\Lambda = 0$. Then Eq. (12.54) shows that, as R gets smaller, the matter term gets more and more important compared to the curvature term $-k/2$. Again this is because $R^2\rho_m$ is proportional either to R^{-1} for matter-dominated dynamics or, even more extremely, to R^{-2} for the radiation-dominated dynamics of the very early universe. Therefore, since our universe is expanding now, it could not have been at rest with $\dot{R} = 0$ at any time in the past. The existence of a Big Bang, *i.e.* whether we reach $R = 0$ at a finite time in the past, depends only on the behavior of the matter; the curvature term is not important, and all three kinds of universes have qualitatively similar histories.

Let us do the computation for a universe that is radiation-dominated, as ours will have been at an early enough time, and that has $\Lambda = 0$ for simplicity. We write $\rho = BR^{-4}$ for some constant B, and we neglect k in Eq. (12.54). This gives

$$\dot{R}^2 = \tfrac{8}{3}\pi BR^{-2},$$

or

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \left(\tfrac{8}{3}\pi B\right)^{1/2} R^{-1}. \tag{12.56}$$

This has the solution

$$R^2 = \left(\tfrac{32}{3}\pi B\right)^{1/2} (t - T), \tag{12.57}$$

where T is a constant of integration. So, indeed, $R = 0$ was achieved at a finite time in the past, and we conventionally adjust our zero of time so that $R = 0$ at $t = 0$, which means we redefine t so that $T = 0$.

Note that we have found that a radiation-dominated cosmology with no cosmological constant has an expansion rate where $R(t) \propto t^{1/2}$. If we had done this computation for a matter-dominated cosmology with $\rho_\Lambda = 0$, we would have found $R(t) \propto t^{2/3}$ (see Exer. 19, § 12.6).

What happens if there is a cosmological constant? If the dark energy is positive, there is no qualitative change in the conclusion, since the term involving ρ_Λ simply increases the value of $\dot{R}$ at any value of R, and this brings the time where $R = 0$ closer to the present epoch. *If the matter density has always been positive, and if the cosmological constant is non-negative, then Einstein's equations make the Big Bang inevitable: the universe began with* $R = 0$ *at a finite time in the past.* This is called the *cosmological singularity*: the curvature tensor is singular, tidal forces become infinitely large, and Einstein's equations do not allow us to continue the solution to earlier times. Within the Einstein framework we cannot ask questions about what came before the Big Bang: time simply began there.

How certain, then, is our conclusion that the universe began with a Big Bang? First, we must ask if isotropy and homogeneity were crucial; the answer is no. The 'singularity theorems' of Penrose and Hawking (see Hawking and Ellis 1973) have shown that our universe certainly had a singularity in its past, regardless of how asymmetric it may have been. But the theorems predict only the *existence* of the singularity: the nature of the singularity is unknown, except that it has the property that at least one particle in the present universe must have originated in it. Nevertheless, the evidence is strong indeed that we *all* originated in it. Another consideration however is that we don't know the laws of physics at the incredibly high densities ($\rho \to \infty$) which existed in the early universe. The singularity theorems of necessity assume (1) something about the nature of $T^{\mu\nu}$, and (2) that Einstein's equations (without cosmological constant) are valid at all R.

The assumption about the positivity of the energy density of matter can be challenged if we allow quantum effects. As we saw in our discussion of the Hawking radiation in the previous chapter, fluctuations can create negative energy for short times. In principle, therefore, our conclusions are not reliable if we are within one *Planck time* $t_{\mathrm{Pl}} = GM_{\mathrm{Pl}}/c^2 \sim 10^{-43}$ s of the Big Bang! (Recall the definition of the Planck mass in Eq. (11.111).) This is the domain of quantum gravity, and it may well turn out that, when we have a quantum theory of the gravitational interaction, we will find that the universe has a history before what we call the Big Bang.

Philosophically satisfying as this might be, it has little practical relevance to the universe we see today. We might not be able to start our universe model evolving from $t = 0$, but we can certainly start it from, say, $t = 100t_{\mathrm{Pl}}$ within the Einstein framework. The primary uncertainties about understanding the physical cosmology that we see around us are, as we will discuss below, to be found in the physics of the early universe, not in the time immediately around the Big Bang.

So far we have restricted our attention to the case of a positive cosmological constant. While this seems to be the most relevant to the evolution of our universe, cosmologies with negative cosmological constant are also interesting. We leave their exploration to the exercises.

Einstein introduced the cosmological constant in order to allow his equations to have a static solution, $\dot{R} = 0$. He did not know about the Hubble flow at the time, and he followed the standard assumption of astronomers of his day that the universe was static. Even in the framework of Newtonian gravity, this would have presented problems, but no-one seems to have tried to find a solution until Einstein addressed the issue within general relativity. We have to do more than just set $\dot{R} = 0$ in Eq. (12.54); we have to guarantee that the solution is an equilibrium one, that the dynamics won't change $\dot{R}$, *i.e.* that the universe is at a minimum or maximum of the 'potential' we discussed earlier. We show in Exer. 20, § 12.6 that the static solution requires

$$\rho_\Lambda = \frac{1}{2}\rho_0.$$

For Einstein's static solution, the dark energy density has to be exactly half of the matter energy density. We shall see below that in our universe the measured value of the dark energy density is about twice that of the matter energy density, so we are near to but not exactly at Einstein's static solution.

Critical density and the parameters of our universe

If we divide Eq. (12.54) by $4\pi R^2/3$, we obtain a version that is instructive for discussions of the physics of the universe:

$$\frac{3H^2}{8\pi} = -\frac{3k}{8\pi R^2} + \rho_m + \rho_\Lambda, \tag{12.58}$$

where we have substituted the Hubble parameter H for $\dot{R}/R$. Since the last two terms on the right are energy densities, it is useful to interpret the other terms in that way. Thus, the Hubble expansion has associated with it an energy density $\rho_H = 3H^2/8\pi$, and the spatial curvature parameter contributes an effective energy density $\rho_k = -3k/8\pi R^2$. This equation becomes

$$\rho_H = \rho_k + \rho_m + \rho_\Lambda.$$

Now, if in the universe today the 'physical' energy density $\rho_m + \rho_\Lambda$ is less than the Hubble energy density ρ_H, then (as we have seen before), the curvature energy density must be positive, the curvature parameter k must be negative, and the universe has hyperbolic hypersurfaces. Conversely, if the physical energy density is larger than the Hubble energy density, the universe will be the closed model. The Hubble energy density is therefore a threshold, and we call it the *critical energy density* ρ_c:

$$\rho_c = \frac{3}{8\pi}H_0^2. \tag{12.59}$$

The ratio of any energy density to the critical is called Ω with an appropriate subscript. Thus, we can divide the earlier energy-density equation, evaluated at the present time, by ρ_c to get

$$1 = \Omega_k + \Omega_m + \Omega_\Lambda. \tag{12.60}$$

These are the quantities used to label the curves in Fig. 12.3. The data from supernovae, the cosmic microwave background, and studies of the evolution of galaxy clusters (below) all suggest that our universe at present has

$$\Omega_\Lambda = 0.7, \Omega_m = 0.3, \Omega_k = 0. \tag{12.61}$$

These mean that we live in a flat universe, dominated by a positive cosmological constant.

What size do these numbers have? It is conventional among astronomers to normalize the Hubble constant H_0 to the value 100 km s^{-1} Mpc^{-1} by introducing the scaled Hubble constant h (nothing to do with gravitational wave amplitudes!):

$$h = H_0/100\,\mathrm{km\,s^{-1}\,Mpc^{-1}}. \tag{12.62}$$

The best value today is $h = 0.71$. Using this, the critical energy density is

$$\rho_c = 1.88 \times 10^{-26} h^2 \mathrm{kg\,m^{-3}} = 9.5 \times 10^{-27} \mathrm{kg\,m^{-3}}.$$

As we have noted, the matter energy density is about 0.3 times this, and this is much more than astronomers can account for by counting stars and galaxies. In fact, studies of the formation of elements in the early universe (below) tell us that the density of *baryonic* matter (normal matter made of protons, neutrons, and electrons) has $\Omega_b = 0.04$. So most of the matter in the universe is non-baryonic, does not emit light, and can be studied astronomically only indirectly, through its gravitational effects. This is called *dark matter*. So we can split Ω_m into its components:

$$\Omega_m = \Omega_b + \Omega_d, \quad \Omega_b = 0.04, \quad \Omega_d = 0.26. \tag{12.63}$$

We will return in § 12.4 below to a discussion of the nature and distribution of the dark matter. The values in Eqs. (12.61) and (12.63) are commonly referred to as the *concordance cosmology*.

The variety of possible cosmological evolutions and the data are captured in the diagram in Fig. 12.4. The evidence is getting rather strong that the dark energy is present, and even dominant. That raises new, important questions. The deepest is, where in physics does this energy come from? We will mention below some of the speculations, but at present there is simply no good theory for it. In such a situation, better data might help. For example, astronomers could try to determine if the dark energy density really is constant in time (as it would be if it comes from a cosmological constant) or variable, which would indicate that it comes from some physical field masquerading as a cosmological constant. As of this writing, new space and ground-based observing programs are being planned, so that in another decade we might have a new generation of ultra-precise measurements of the dark energy.

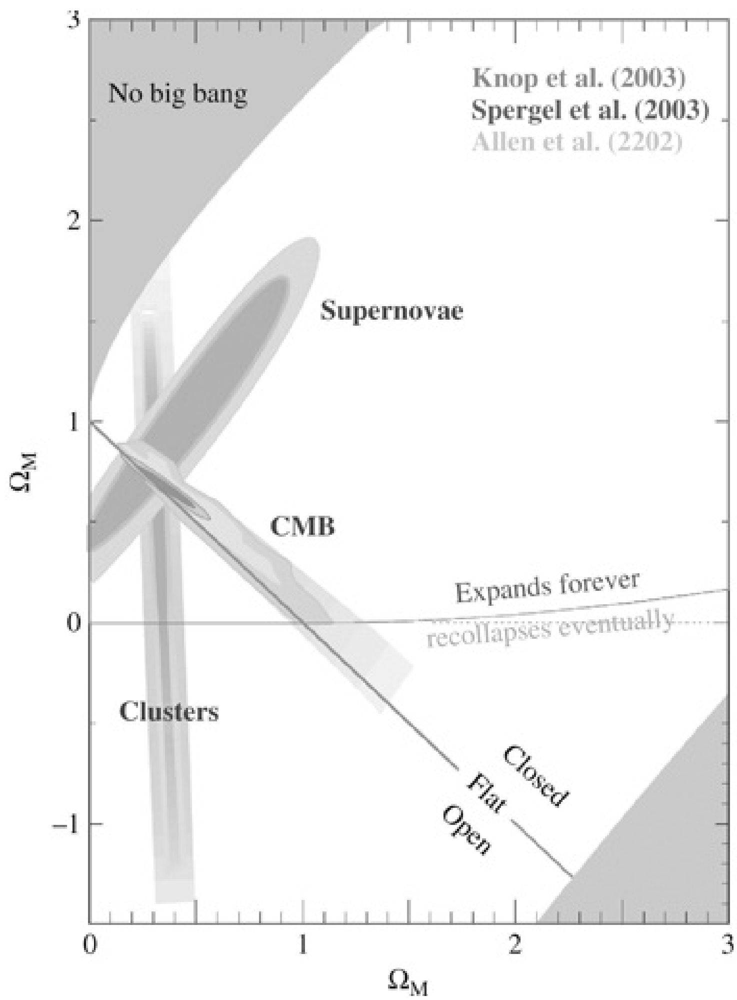


Figure 12.4 In the $\mathbf{\Omega}_m$ v. $\mathbf{\Omega}_\Lambda$ plane one sees the variety of possible cosmological models, their histories and futures. The constraints from studies of supernovae (Knop *et al*. 2003), the cosmic microwave background radiation (Spergel *et al*.

2003), and galaxy clustering (Allen *et al*. 2002) are consistent with one another and all overlap in a small region of parameter space centered on $\mathbf{\Omega}_\Lambda = 0.7$ and $\mathbf{\Omega}_m = 0.3$. This means that $\mathbf{\Omega}_k = 0$ to within the errors. Figure courtesy the Supernova Cosmology Project.

From the point of view of general relativity, one of the most intriguing ways of studying the dark energy is with the LISA gravitational wave detector. As mentioned in § 9.5, LISA will be able to observe coalescences of black holes at high redshifts and measure their distances. What will be measured from the signal is the luminosity distance d_L to the binary, since it is based on an inference of the luminosity of the system in gravitational waves from the information contained in the signal. This measurement can be made with great accuracy, perhaps with errors at the few percent level. To do cosmography, we have to combine these luminosity distance measures with redshifts, and that will not be easy: black hole coalescences do not give off any electromagnetic radiation directly, so it will not be easy to identify the galaxy in which the event has occurred. But the galaxies hosting the mergers will not be normal galaxies, and the merger event might be accompanied by other signs, such as an alteration of X-ray luminosity, the existence of unusual jets, a disturbed morphology. It may well be possible in many cases to identify the host galaxy within the LISA position error box, which may be less than 10 arcminutes in size in favorable cases.

A gravitational-wave measurement would be a very desirable complement to other studies of the dark energy, because it needs no calibration: it would be independent of the assumptions of the cosmic distance ladder. It would therefore be an important check on the systematic errors of other methods.

12.4 Physical cosmology: the evolution of the universe we observe

The observations described in the last section confirm the reliability of using a general-relativistic cosmological model with dark energy to describe the evolution of the universe, starting as far back as our observations can take us. During the last few decades, astrophysicists have developed a deep and rich understanding of how the universe we see, with all its structure and variety, evolved out of a homogeneous hot expanding plasma. The story is a fascinating one that we can only sketch here. But it is fair to say that there is now a consistent story that goes from the moment that protons and neutrons became identifiable particles right up to the formation of stars like our Sun and planets like our Earth. Many of the details are poorly understood, especially where observations are difficult to perform, but the physical framework for understanding them is not in doubt.

The expansion of the universe was accompanied by a general cooling off of its matter: photons have been redshifted, the random velocities of gas particles dropped, structures like galaxies and stars condensed out. The history of the universe is therefore a thermal history: instead of using cosmological time t or the scale factor R to mark different stages of evolution, we will use temperature, or equivalently energy, converting between them by $E = kT$. That brings us closer to the physics.

Our understanding of the history of the universe rests on our understanding of its physical laws, and these are tested up to energies of order 1 TeV in modern particle colliders. So our physical picture of the evolution of the universe can reliably start when the expanding plasma had that sort of energy.

Decoupling: forming the cosmic microwave background radiation

If we start at the present moment and go backwards, the matter energy density increases as the scale factor R decreases, but the dark energy density remains constant, so (unless the dark energy comes from some exotic physics) we can safely ignore the dark energy at early times. The density of ordinary matter (dark and baryonic) increases as R^{-3}, while the energy density of the photons of the cosmic microwave background increases as R^{-4}. Since the energy density of the cosmic microwave background today is $\Omega_\gamma \sim 10^{-5}$ and the matter density is $\Omega_m = 0.3$, they will have been equal when the scale factor was a factor of 3×10^4 smaller than today. Since redshift and scale factor go together, this is the redshift when the expanding universe changed from radiation-dominated to matter dominated. At higher redshifts, the universe was radiation-dominated. The temperature was about 10^5K and the energy scale was about 10 eV. This happened about 3000 years after the Big Bang.

Now, this energy is near the ionization energy of hydrogen, which is 13.6 eV. This is an important number because hydrogen is the principal constituent of the baryonic matter. If the temperature is high enough to ionize hydrogen, the universe will be filled with a plasma that is opaque to electromagnetic radiation. Once hydrogen cools off enough to become neutral, the remaining photons in the universe will be able to move through it with a low probability of scattering. This moment of *decoupling* (also called *recombination*) defines the moment at which the cosmic microwave background radiation was created. This actually occurs at a rather smaller energy than 13.6 eV, since there is enough hydrogen to stop the photons even when only a small portion of it is ionized. The epoch of decoupling occurred at a temperature a bit below 1 eV, at a time when the universe was matter-dominated. The redshift was about 2000, and the time was about 4×10^5 years after the Big Bang.

Observations of the cosmic microwave background reveal that it has an almost perfect black-body spectrum with a temperature of T = 2.725K. But they also show that it has small but significant temperature irregularities, departures from strict homogeneity that are the harbingers of the formation of galaxy clusters and galaxies. A map of these is shown in Fig. 12.5. The temperature irregularities are of order 10^{-5} of the background temperature, and they are caused by irregularities in the matter distribution of the same relative size. Small as these may seem, numerical studies show that they are adequate to lead to all the structure we see today. Because the dominant form of matter is the dark matter, the density fluctuations that are seen in Fig. 12.5 and that led to galaxy formation were in its distribution. Simulations show that only if the random velocities of dark matter particles were small could the small irregularities grow in size fast enough to trap baryonic gas and make it form galaxies. The matter had, therefore, to be *cold*, and we call this model of galaxy formation the *cold dark matter model*. The standard cosmological model is called ΛCDM: cold dark matter with a cosmological constant.

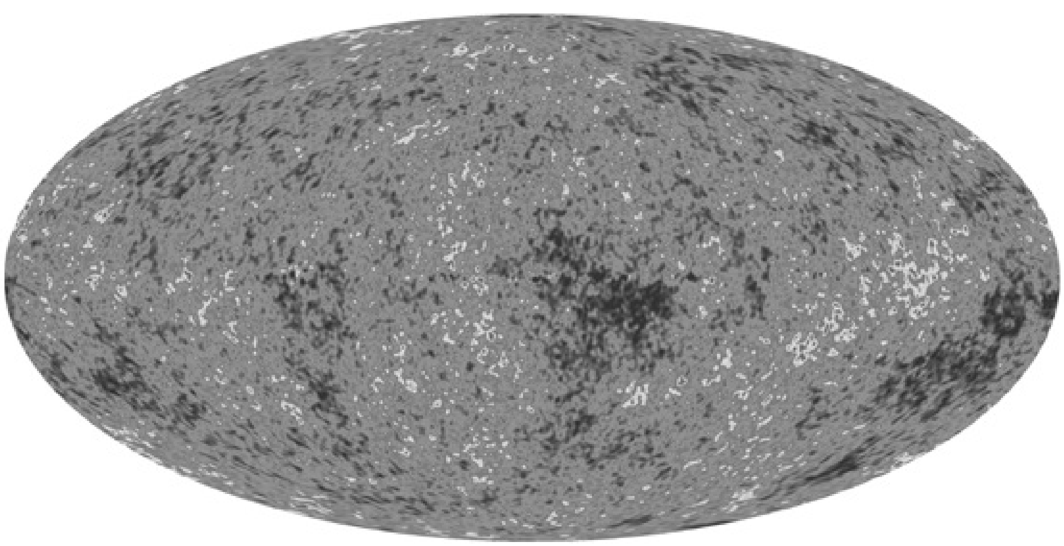


Figure 12.5 A map of the small-scale temperature inhomogeneities of the cosmic microwave background, made by the WMAP satellite (Spergel, *et al*,

2003). The range of fluctuations is ±200μK. Figure courtesy the WMAP project and NASA.

The parameters of our cosmology – the cosmological constant, matter fraction, and so on – leave their imprint on these fluctuations. The fluctuations occur on all length-scales, but they do not have the same size on different scales. The angular spectrum of fluctuations contains a rich amount of information about the cosmological parameters, and it is here that we find the constraints shown in Fig. 12.4.

Dark matter and galaxy formation: the universe after decoupling

Going forward from the time of decoupling, physicists have simulated the evolution of galaxies and clusters of galaxies from the initial perturbations. The density perturbations in the dark matter grow slowly, as they have been doing since before decoupling. But before decoupling, the baryonic matter could not respond very much to them, because it remained in equilibrium with the photons. Once the baryonic matter took the form of neutral atoms, it could begin to fall into the gravitational wells created by the dark matter.

Unlike the dark matter, the baryonic matter had the ability to concentrate itself at the bottoms of these wells, so that the density irregularities of the baryonic matter soon became stronger than those of the dark matter. The reason for this is that the baryonic matter was charged: as the atoms fell into the potential wells, they collided with one another and collisionally excited their electrons into higher energy levels. The density was low enough that the electrons then decayed back to the ground state by radiating away the excess energy. Now that decoupling had happened, this was a one-way street, a way of extracting energy from the baryonic gas and allowing it to clump inside the dark-matter wells. Astronomers call this process *cooling*, even though the net effect of radiating energy away is to make the baryonic matter hotter as it falls deeper into the potential wells!

The dark matter itself is not charged, so it cannot form such strong contrast. It forms extended 'halos' around galaxies today, as we shall see below. Extensive numerical simulations using supercomputers show that the clumps of baryonic gas eventually began to condense into basic building-block clumps of a million solar masses or so, and then these began to merge together to form galaxies. So although images of the universe seem to show a lot of well-separated galaxies, the fact is that most of the objects we see were formed from many hundreds or more of mergers. Mergers are still going on: astronomers have discovered a fragment of several million solar masses that is currently being integrated into our own Milky Way galaxy, on the other side of the center from our location. The unusual star cluster Omega Centauri seems to be the core of such a mini-galaxy that was absorbed by the Milky Way long ago. As mentioned in § 11.4, astronomers have found a massive black hole in its center. And the Magellanic Clouds, easily visible in the sky in the southern hemisphere, may be on their way to merging into the Milky Way.

In 2012 the European Space Agency plans to launch the astrometry satellite GAIA, which will measure the positions and proper motions of a billion stars to unprecedented accuracy. One of the many goals of GAIA will be to map the velocity field of the Milky Way, because astronomers expect that ancient mergers will still be apparent as ‘streams’ of stars moving differently from others.

Sometime during this hierarchy of merging structures, the density of the gas got high enough for the first generation of stars to form. These are called Population III stars, and they were unlike anything we see today. Since the gas from which they formed was composed only of hydrogen and helium, with none of the heavier elements that were made by this generation of stars and incorporated into the next, these stars were much more massive. With masses between 100 and 1000 $M_\odot$, they became very hot, evolved quickly, generated heavier elements, blew much of their outer layers and much of the new elements away with strong stellar winds, and then very likely left behind a large population of black holes. The ultraviolet light emitted by these stars seems to have re-ionized much of the hydrogen in the universe, which had been neutral since decoupling. All of this happened between redshifts of 10 and 20, the epoch of re-ionization. This was the epoch of first light for galaxies, the first time that the expanding universe would have looked optically a bit like it does today, if there had been anyone there to observe it!

After re-ionization and the generation of heavier elements by the first stars, the continued expansion of the universe led galaxies to be more and more isolated from each other, and they became nurseries for one generation of new stars after another, each with a bit more of the heavier elements. Our Sun, whose age is about 5 billion years, was formed at a redshift between 0.3 and 0.4.

One of the puzzles in this scenario of hierarchical structure formation is the appearance of massive black holes in the centers of apparently all galaxies. Astrophysicists do not yet know whether these formed directly by the collapse of huge gas clouds as the baryonic matter was accumulating in the potential wells, or if they arose later by the growth and merger of intermediate-mass black holes left behind by Population III stars. The fact that galaxy evolution is dominated by mergers suggests that the LISA gravitational wave observatory (see the previous chapter) will have an abundance of black-hole mergers to study.

Although physicists do not know what kinds of particles (or indeed, more massive structures like black holes) might make up the dark matter, and although the dark matter emits no electromagnetic radiation, it is possible to make indirect observations of it. One way it shows itself is in the rotation curves of spiral galaxies: in the outer regions of spirals the orbital speeds of gas and stars are much greater than could be accounted for by the gravitational pull of the visible stars. Indeed, observations suggest that in most spiral galaxies the total mass of the galaxy inside a given radius continues to increase linearly with radius even well outside the visible limits of the galaxy. This is the footprint of the dark matter density concentration within which the galaxy formed. The other way of getting indirect evidence for dark matter is through gravitational lensing. Astronomers have many images like Fig. 11.8, as we discussed in the previous chapter. Both of these methods involve essentially using gravity to 'weigh' the dark matter.

The early universe: fundamental physics meets cosmology

If we go back in time from the moment when the radiation and matter densities were equal, then we are in a radiation-dominated universe. Eventually, when we are only about about 200s away from the Big Bang, the temperature rises to about 50 keV, the mass difference between a neutron and a proton. This is the temperature at which nuclear reactions among protons and neutrons come into equilibrium with each other. Above this energy all the baryons were free. As the universe cooled through this temperature, some heavier elements were formed: mainly ^{4}He, but also small amounts of ^{3}He, Li, B, and traces of other light elements. All the lithium and helium we see in the universe today was formed at this time: processes inside stars tend to destroy light elements, not make them. The final abundances of these elements is very sensitive to the rate at which the universe was expanding at this time. From extensive computations of the reaction networks, astrophysicists have been able to show that the universe contains no significant amounts of light or massless particles other than photons and the three types of neutrinos that are known from particle-physics experiments. If there were others, their self-gravity would have slowed the universe more strongly, which means that to match the Hubble expansion today, a universe with such extra particles would have had to have been expanding faster than the nucleosynthesis computations allow. If there are extra particles, they must have an energy density today that is significantly less than that of the photons, which have $\Omega_\gamma \sim 10^{-5}$. Gravitational waves from the Big Bang must, therefore, satisfy this nucleosynthesis bound.

Notice that we are already within 200 s of the Big Bang in this discussion, and still we are in the domain of well-understood physics. At about 1 s, the temperature was around 500 keV, which is the mass of the electron. In this plasma, therefore, there was an abundance of electrons and positrons, constantly annihilating against one another and being created again by photons. Much earlier than this the rest mass of the electrons is negligible, so the number and energy density of photons and of electrons and positrons was similar. As the universe expanded through this 500 keV temperature and cooled, the electrons and positrons continued to annihilate, but no more were produced. After a few seconds, there were apparently essentially no positrons, and there was about one electron for every 10^9 photons. This ratio of 10^9 is called the specific entropy of the universe, a measure of its disorder.

Why were there any electrons left at all after this annihilation phase? Why, in other words, was there any matter left over to build into planets and people? Extensive observational programs, coupled to numerical simulations, have convincingly established that there is no ‘missing’ antimatter hidden somewhere, no anti-stars or anti-galaxies: significant amounts of antimatter just do not exist any more. Clearly, during the equilibrium plasma phase, electrons and positrons were not produced in equal numbers. The same must also have happened at a much earlier time, when protons and antiprotons were in equilibrium with the photon gas, when the temperature was above a few hundred MeV (only 10 μs after the Big Bang): something must have favored protons over antiprotons in the same ratio as for electrons over positrons, so that the overall plasma remained charge-neutral. This is one of the central mysteries of particle physics. Something in the fundamental laws of physics gave a slight preference to electrons. *Nature has a mattter-antimatter asymmetry.*

At times earlier than 10^{-5}s, there are no protons or neutrons, just a plasma of quarks and gluons, the fundamental building blocks of baryons. According to particle theory, quarks are ‘confined’, so that we never see a free one detached from a baryon. But at high enough temperatures and densities, the protons overlap so much that the quarks can stay confined and still behave like free particles.

We can even push our perspective another step higher in temperature, to around 10 TeV, which is the frontier for current accelerators. Physics is not well understood at these energies, and the Large Hadron Collider at CERN in Geneva will soon (end of 2008) start doing experiments to look for the Higgs particle and to find evidence for supersymmetry. Both of these are theoretical constructs designed to solve deep theoretical problems in fundamental physics. In particular, supersymmetry has the advantage of making it easier within particle physics theories to predict a value of the cosmological constant in the range of what we observe. If supersymmetry is found, it will encourage the idea that the dark energy can be accommodated within the current framework of fundamental physics theory. At 10 TeV, we are just 10^{-14}s after the Big Bang. Although physics is poorly known, it is unlikely that anything happened at this point that will challenge the picture presented here of what happened later.

In this scheme there is one important thing that is missing: we have mentioned no mechanism in any of this physics for generating the density irregularities in the dark matter that led to galaxy formation. We observe them in the microwave background, and we know that they are needed in order to trigger all the processes that eventually led to our own evolution. But even at 10^{-14}s, they are simply an initial condition: they have to be there, at a much smaller amplitude than we see them in the microwave background because the irregularities grow as the universe expands, but they must be there. And known physics has no explanation for how they got there. The exciting answer to this problem lies in the scenario of inflation, which we come back to below.

But the density perturbations are not the only feature of the early universe that is not explained. Right from the start we have assumed homogeneity and isotropy, based on observations. We have had to accept a small amount of inhomogeneity in the density distribution at the time of decoupling, but that was inevitable: our assumption of homogeneity does not hold on small scales, where galaxies and planets form. On the large scale, we need to ask, why is the universe so smooth? This is particularly difficult to explain because in the standard Big Bang, there is no physical process that could work to smooth things out. This needs some explanation.

Consider the primordial abundance of helium. It was fixed when the universe was only five minutes old. When we look with our telescopes in opposite directions on the sky, we can see distant quasars and galaxies that appear to have the same element abundances as we do, and yet they are so far away from each other that they could not have been in communication: they are outside each other's particle horizon in the standard cosmological model. One way to 'explain' this is simply to postulate that the initial conditions for the Big Bang were the same everywhere, even in causally disconnected regions. But it would be more satisfying physically if some process could be found that enabled these regions to communicate with each other at a very early time, even though they appear to be disconnected. Again, inflation offers such a mechanism. *Inflation.* The basic idea of inflation (Starobinsky 1980, Guth 1981, Linde 1982) is that, at a very early time like 10^{-35}s, the universe was dominated by a large positive cosmological constant, much larger than we have today, but one that was only temporary: it turned on at some point and then turned off again, for reasons we will discuss below. But during the time when the universe was dominated by this constant, the matter and curvature were unimportant, and the universe expanded according to the simple law

$$H^2 = \tfrac{8}{3}\pi\rho_\Lambda \qquad \Longrightarrow \qquad \frac{\dot{R}}{R} = 1/\tau_\Lambda, \tag{12.64}$$

which is an exponential law with a growth time

$$\tau_\Lambda = \left(\frac{3}{8\pi\rho_\Lambda}\right)^{1/2}. \tag{12.65}$$

If this exponential expansion lasted 20 or 30 e-foldings, then a region of very small size could have been inflated into the the size of a patch that would be big enough to become the entire observable universe today. The idea is that, before inflation, this small region had been smoothed out by some physical process, which was possible because it was small enough to do this even in the time available. Then inflation set in and expanded it into the initial data for our universe.

This would explain the homogeneity of what we see: everything did indeed come from the same patch. And inflation also explains the fluctuation in the cosmic microwave background. Here we have to go into more detail about the mechanism for inflation. Attempts to compute the cosmological constant today focus on the vacuum energy of quantum fields, which we used in order to explain the Hawking radiation in the previous chapter. The vacuum energy is attractive for this purpose because the vacuum must be invariant under Lorentz transformations: there should not be any preferred observer for empty space in quantum theory. This means that any stress-energy tensor associated with the vacuum must be Lorentz invariant. Now, the only Lorentz-invariant symmetric tensor field of type $\binom{0}{2}$ is the metric tensor itself, so any vacuum-energy explanation of dark energy will automatically produce something like a cosmological constant, proportional to the metric tensor.

In some models of the behavior of the physical interactions at very high energies, beyond the TeV scale, it is postulated that there is a phase transition in which the nature of the vacuum changes, and a large amount of vacuum energy is released in the form of a cosmological constant, powering inflation. But this is a dynamical process, which sets in when the phase change occurs and then stops when the energy is converted into the real energy that eventually becomes the particles and photons in our universe. So for a limited time, the universe inflates rapidly. Now, at the beginning there are the usual vacuum fluctuations, and the remarkable thing is that the exponential inflation amplifies these fluctuations in much the same way as a nonlinear oscillator can pump up its oscillation amplitude. When inflation finishes, what were small density perturbations on the quantum scale have become much larger, classical perturbations.

When physicists perform computations with this model, it does very well. The amplitude of the fluctuations is reasonable, their spectrum matches that which is inferred from the cosmic microwave background, and the physical assumptions are consistent with modern views of unification among the various interactions of fundamental physics. Inflation will remain a ‘model’ and not a ‘theory’ until either a full theory unifying the nuclear forces is found or until some key observation reveals the fields and potential that are postulated within the model. However, it is a powerful and convincing paradigm, and it is currently the principal framework within which physicists address the deepest questions about the early universe.

Beyond general relativity

Inflation goes beyond standard physics, making assumptions about the way that the laws governing the nuclear interactions among particles behave at the very high energies that obtained in the early universe. But it does not modify gravity: it works within classical general relativity. Nevertheless, as we have remarked before, the classical theory must eventually be replaced by a quantum description of gravitation, and the search for this theory is a major activity in theoretical physics today.

Although no consistent theory has yet emerged, the search has produced a number of exciting ideas that offer the possibility of new kinds of observations, new kinds of explanations. One approach, called loop quantum gravity, directly attacks the problem of how to quantize spacetime, ignoring at first the other forces in spacetime, like electromagnetism. On a fine scale, presumably the Planck scale, it postulates that the manifold nature of spacetime breaks down, and the smaller-scale structure is one of nested, tangled loops. There are a number of variants on this approach, with different structures, but the common idea is that spacetime is a coarse-grained average over something that has a much richer topology. These ideas come from the mathematics in a natural way. A recent triumph of loop quantum gravity is to show that the Big Bang may not have been singular after all, that going backwards in time the universe is able to pass through the Big Bang and become a classical collapsing universe on the other side (Bojowald 2005).

Even more active, in terms of the number of physicists working in it, is the string-theory approach to quantum gravity. Here the aim is to unify all the interactions, including gravity, so the theory includes the nuclear and electromagnetic interactions from the start. String theory seems to be consistent, in the sense of not having to do artificial things to get rid of infinite energies, only in 11 spacetime dimensions. We live in just four of these, so physicists are beginning to ask questions about the remaining ones.

The first assumption was that they never got big: that attached to each point is a Plancksized seven-sphere offering the possibility of exiting from our four-dimensional universe only to things that are smaller than the Planck length. This would not be easy to observe. But it is also possible that some of these extra dimensions are big, and our four-dimensional universe is simply a four-surface in this five-or more-dimensional surrounding. This surface has come to be called a *brane*, from the word ‘membrane’. String theory on branes has a special property: electromagnetism and the nuclear forces are confined to our brane, but gravitation can act in the extra dimensions too. This would lead to a modification of the inverse-square-law of Newtonian gravity on short distances, on scales comparable to a relevant length-scale in the surrounding space. All we can say is that this scale must be smaller than about a millimeter, from experiments on the inverse square law. But there are many decades between a millimeter and the Planck scale, and new physics might be waiting to be discovered anywhere in between (Maartens 2004).

The new physics could take many different forms. Some kind of collision with another brane might have triggered the Big Bang. A nearby extra brane might have a parallel world of stars and galaxies, interacting with us only through gravity: shadow matter. There might be extra amounts of gravitational radiation, due either to shadow matter or to unusual brane-related initial conditions at the Big Bang.

Although these ideas sound like science fiction, they are firmly grounded in model theories, which are deliberate over-simplifications of the full equations of string theory, and which involve deliberate choices of the values of certain constants in order to get these strange effects. They should be treated as neither predictions nor idle speculation, but rather as harbingers of the kind of revolutionary physics that a full quantum theory of gravity might bring us. Experimental hints, from high-precision physics, or observational results, perhaps from gravitational waves or from cosmology, might at any time provide key clues that could point the way to the right theory.

12.5 Further reading

The literature on cosmology is vast. In the body of the chapter I have given the principal references to original results, so I list here some recommended books on the subject.

Standard cosmology is treated in great detail in Weinberg (1972). Cosmological models in general relativity become somewhat more complex when the assumption of isotropy is dropped, but they retain the same overall features: the Big Bang, open vs. closed. See Ryan and Shepley (1975). A well-balanced introduction to cosmology is Heidmann (1980).

Early discussions on physical cosmology that remain classics include Peebles (1980), Liang and Sachs (1980), and Balian *et al.* (1980). More modern is Liddle (2003).

An important current research area is into inhomogeneous cosmologies. See

MacCallum (1979). Another subject closely allied to theoretical cosmology is singularity theory: Geroch and Horowitz (1979), Tipler *et al.* (1980). See also the stimulating article by Penrose (1979) on time asymmetry in cosmology.

For greater depth on physical cosmology, see the excellent text by Mukhanov (2005). For a different point of view on 'why' the universe has the properties it does, see the book by Barrow and Tipler (1986) on the anthropic principle.

For popular-level cosmology articles written by research scientists, see the Einstein Online website: `http://www.einstein-online.info/en/`.

12.6 Exercises

Use the metric of a two-sphere to prove the statement associated with Fig. 12.1, that the rate of increase of the distance between any two points as the sphere expands (as measured *on* the sphere!) is proportional to the distance between them.

The astronomer's distance unit, the parsec, is defined to be the distance from the Sun to a star whose parallax is exactly one second of arc. (The parallax of a star is half the maximum change in its angular position as measured from Earth as Earth orbits the Sun.) Given that the radius of Earth's orbit is 1 AU = 10^{11} m, calculate the length of one parsec.

Newtonian cosmology.

(a) Apply Newton's law of gravity to the study of cosmology by showing that the general solution of $\nabla^2\Phi = 4\pi\rho$ for ρ = const. is a quadratic polynomial in Cartesian coordintes, but is not necessarily isotropic.

(b) Show that if the universe consists of a region where ρ = const., outside of which there is vacuum, then, if the boundary is not spherical, the field will not be isotropic: the field will show significant deviations from sphericity throughout the interior, even at the center.

(c) Show that, in such a Newtonian cosmology, an experiment done locally could determine the shape of the boundary, even if the boundary is far outside our particle horizon.

Show that if $h_{ij}(t_1) \neq f(t_1, t_0)h_{ij}(t_0)$ for all i and j in Eq. (12.3), then distances between galaxies would increase anisotropically: the Hubble law would have to be written as

$$v^i = H_j{}^i x^j \tag{12.66}$$

for a matrix $H^i{}_j$ not proportional to the identity.

Show that if galaxies are assumed to move along the lines x^i = const., and if we see the local universe as homogeneous, then g_{0i} in Eq. (12.5) must vanish.

a) Prove the statement leading to Eq. (12.8), that we can deduce G_{ij} of our three-spaces by setting Φ to zero in Eqs. (10.15)–(10.17).

(b) Derive Eq. (12.9).

Show that the metric, Eq. (12.7), is not locally flat at $r = 0$ unless $A = 0$ in Eq. (12.11).

a) Find the coordinate transformation leading to Eq. (12.19).

(b) Show that the intrinsic geometry of a hyperbola $t^2 - x^2 - y^2 - z^2$ = const. > 0 in Minkowski spacetime is identical with that of Eq. (12.19) in appropriate coordinates.

(c) Use the Lorentz transformations of Minkowski space to prove that the $k = -1$ universe is homogeneous and isotropic.

a) Show that a photon which propagates on a radial null geodesic of the metric, Eq. (12.13), has energy $-p_0$ inversely proportional to $R(t)$.

(b) Show from this that a photon emitted at time t_e and received at time t_r by observers at rest in the cosmological reference frame is redshifted by

$$1 + z = R(t_r)/R(t_e). \tag{12.67}$$

Show from Eq. (12.24) that the relationship between velocity and cosmological redshift for a nearby object (small light-travel-time to us) is $z = v$, as we would expect for an object with a recessional velocity v.

) Prove Eq. (12.29) and deduce Eq. (12.30) from it.

(b) Fill in the indicated steps leading to Eq. (12.31).

Derive Eq. (12.42) from Eq. (12.31). Derive Eq. (12.44) from Eq. (12.40).

Astronomers usually do not speak in terms of intrinsic luminosity and flux. Rather, they use absolute and apparent magnitude. The (bolometric) *apparent magnitude* of a star is defined by its flux F relative to a standard flux F_s:

$$m = -2.5 \log_{10}(F/F_s) \tag{12.68}$$

where $F_s = 3 \times 10^{-8}$ J m^{-2} s^{-1} is roughly the flux of visible light at Earth from the brightest stars in the night sky. The *absolute magnitude* is defined as the apparent magnitude the object would have at a distance of 10 pc:

$$M = -2.5 \log_{10}[L/4\pi(10\,\text{pc})^2 F_s]. \tag{12.69}$$

Using Eq. (12.42), with Eq. (12.27), rewrite Eq. (12.34) in astronomer's language as:

$$m - M = 5\log_{10}(z/10\,\mathrm{pc}\,H_0) + 1.09(1-q_0)z. \tag{12.70}$$

Astronomers call this the *redshift-magnitude relation*.

a) For the Robertson–Walker metric Eq. (12.13), compute all the Christoffel symbols $\Gamma^{\mu}{}_{\alpha\beta}$. In particular show that the nonvanishing ones are:

$$\begin{aligned}
&\Gamma^0{}_{jk} = \frac{\dot R}{R} g_{jk}, \quad \Gamma^j{}_{0k} = \frac{\dot R}{R}\delta^j{}_k, \quad \Gamma^r{}_{rr} = \frac{kr}{1-kr^2},\\
&\Gamma^r{}_{\theta\theta} = -r(1-kr^2), \quad \Gamma^r{}_{\phi\phi} = -r(1-kr^2)\sin^2\theta,\\
&\Gamma^\theta{}_{r\theta} = \Gamma^\phi{}_{r\phi} = \frac{1}{r}, \quad \Gamma^\theta{}_{\phi\phi} = \sin\theta\,\cos\theta, \quad \Gamma^\phi{}_{\theta\phi} = \cot\theta.
\end{aligned} \tag{12.71}$$

(b) Using these Christoffel symbols, show that the time-component of the divergence of the stress-energy tensor of the cosmological fluid is

$$T^{0\alpha}{}_{;\alpha} = \dot\rho + 3(\rho + p)\frac{\dot R}{R}. \tag{12.72}$$

(c) By multiplying this equation by R^3, derive Eq. (12.46).

Show from Eq. (12.49) that if the radiation has a black-body spectrum of temperature T, then T is inversely proportional to R.

Use the Christoffel symbols computed in Exer. 14 above to derive Eq. (12.50).

Use Eq. (12.46) and the time-derivative of Eq. (12.54) to derive Eq. (12.55) for $\ddot R$. Make sure you use the fact that $p_\Lambda = -\rho_\Lambda$.

In this chapter we saw that the negative pressure (tension) of the cosmological constant is responsible for accelerating the universe. But is this a contradiction to ordinary physics? Does a tension pull inward, not push outward? Resolve this apparent contradiction by showing that the net pressure *force* on any local part of the universe is zero. Refer to the discussion at the end of § 4.6.

Assuming the universe to be matter-dominated and to have zero cosmological constant, show that at times early enough for one to be able to neglect k in Eq. (12.54), the scale factor evolves with time as $R(t) \propto t^{2/3}$.

Assume that the universe is matter dominated and find the value of ρ_Λ that permits the universe to be static.

(a) Because the universe is matter-dominated at the present time, we can take $\rho_m(t) = \rho_0[R_0/R(t)]^3$ where the subscript '0' refers to the static solution we are looking for. Differentiate the 'energy' equation Eq. (12.54) with respect to time to find the dynamical equation governing a matter-dominated universe:

$$\ddot{R} = \tfrac{8}{3}\pi\rho_\Lambda R - \tfrac{4}{3}\pi\rho_0 R_0^3 R^{-2}. \tag{12.73}$$

Set this to zero to find the solution

$$\rho_\Lambda = \tfrac{1}{2}\rho_0.$$

For Einstein's static solution, the cosmological constant energy density has to be half of the matter energy density.

(b) Put our expression for ρ_m into the right-hand-side of Eq. (12.54) to get an energy-like expression which has a derivative that has to vanish for a static solution. Verify that the above condition on ρ_Λ does indeed make the first derivative vanish.

(c) Compute the second derivative of the right-hand-side of Eq. (12.54) with respect to R and show that, at the static solution, it is positive. This means that the 'potential' is a maximum and *Einstein's static solution is unstable.*

Explore the possible futures and histories of an expanding cosmology with *negative* cosmological constant. You may wish to do this graphically, by drawing figures analogous to Fig. 11.1. See also Fig. 12.4.

(Parts of this exercise are suitable only for students who can program a computer.) Construct a more realistic equation of state for the universe as follows.

(a) Assume that, today, the matter density is $\rho_m = m \times 10^{-27}$ kg m^{-3} (where m is of order 1) and that the cosmic radiation has black-body temperature 2.7 K. Find the ratio $\varepsilon = \rho_r/\rho_m$, where ρ_r is the energy density of the radiation. Find the number of photons per baryon, $\sim \varepsilon m_p c^2/kT$.

(b) Find the general form of the energy-conservation equation, $T^{0\mu}{}_{,\mu} = 0$, in terms of $\varepsilon(t)$ and $m(t)$.

(c) Numerically integrate this equation and Eq. (12.54) for $\Lambda = 0$ back in time from the present, assuming $\dot{R}/R = 75$ km s^{-1} Mpc^{-1} today, and assuming there is no exchange of energy between matter and radiation. Do the integration for

m = 0.3, 1.0, and 3.0. Stop the integration when the radiation temperature reaches $E_i/26.7\ k$, where E_i is the ionization energy of hydrogen (13.6 eV). This is roughly the temperature at which there are enough photons to ionize all the hydrogen: there is roughly a fraction 2×10^{-9} photons above energy E_i when $kT = E_i/26.7$, and this is roughly the fraction needed to give one such photon per H atom. For each m, what is the value of $R(t)/R_0$ at that time, where R_0 is the present scale factor? Explain this result. What is the value of t at this epoch?

(d) Determine whether the pressure of the matter is still negligible compared to that of the radiation. (You will need the temperature of the matter, which equals the radiation temperature now because the matter is ionized and therefore strongly coupled to the radiation.)

(e) Integrate the equations backwards in time from the decoupling time, now with the assumption that radiation and matter exchange energy in such a way as to keep their temperatures equal. In each case, how long ago was the time at which $R = 0$, the Big Bang?

Calculate the redshift of decoupling by assuming that the cosmic microwave radiation has temperature 2.7 K today and had the temperature $E_i/20\ k$ at decoupling, where E_i = 13.6 eV is the energy needed to ionize hydrogen (see Exer. 22c).

If Hubble's constant is 75 km s^{-1} Mpc^{-1}, what is the minimum present density for a $k = +1$ universe?

Estimate the times earlier than which our uncertainty about the laws of physics prevents us drawing firm conclusions about cosmology as follows.

(a) Deduce that, in the radiation-dominated early universe, where the curvature term depending on the curvature constant k (0, 1, −1) is negligible, the temperature T behaves as (from now on, k is Boltzmann's constant)

$$T = \beta t^{-1/2}, \quad \beta = (45\,\hbar^3/32\pi^3)^{1/4} k^{-1}.$$

(b) Assuming that our knowledge of particle physics is uncertain for $kT > 10^3$ GeV, find the earliest time t at which we can have confidence in the physics.

(c) Quantum gravity is probably important when a photon has enough energy kT to form a black hole within one wavelength ($\lambda = h/kT$). Show that this gives $kT \sim h^{1/2}$. This is the *Planck temperature*. At what time t is this an important worry?

A universe *could* be homogeneous but anisotropic, if, for instance, it had a large-scale magnetic field which pointed in one direction everywhere and whose magnitude was the same everywhere. On the other hand, an inhomogeneous universe could not be isotropic about every point, since most – if not all – places in the universe would see a sky that is ‘lumpy’ in one direction and not in another.

A

Appendix A **Summary of linear algebra**

For the convenience of the student we collect those aspects of linear algebra that are important in our study. We hope that none of this is new to the reader.

Vector space

A collection of elements $V = \{A, B, \ldots\}$ forms a *vector space* over the real numbers if and only if they obey the following axioms (with a, b real numbers).

) V is an abelian group with operation + ($A + B = B + A \in V$) and identity 0 ($A + 0 = A$).

) Multiplication of vectors by real numbers is an operation which gives vectors and which is:

(i) distributive over vector addition, $a(A + B) = a(A) + a(B)$;
(ii) distributive over real number addition, $(a + b)(A) = a(A) + b(A)$;
(iii) Associative with real number multiplication, $(ab)(A) = a(b(A))$;
(iv) consistent with the real number identity, $1(A) = A$.

This definition could be generalized to vector spaces over complex numbers or over any field, but we shall not need to do so.

A set of vectors $\{A, B, \ldots\}$ is said to be *linearly independent* if and only if there do not exist real numbers $\{a, b, \ldots, f\}$, not all of which are zero, such that

$$aA + bB + \cdots + fF = 0.$$

The dimension of the vector space is the largest number of *linearly independent* vectors we can choose. A basis for the space is any linearly independent set of vectors $\{A_1, \ldots, A_n\}$, where n is the dimension of the space. Since for any B the set $\{B, A_1, \ldots, A_n\}$ is linearly dependent, it follows that B can be written as a linear combination of the basis vectors:

$$B = b_1A_1 + b_2A_2 + \cdots + b_nA_n.$$

The numbers $\{b_1, \ldots, b_n\}$ are called the components of B on $\{A_1, \ldots, A_n\}$.

An *inner product* may be defined on a vector space. It is a rule associating with any pair of vectors, A and B, a real number $A \cdot B$, which has the properties:

(1) $A \cdot B = B \cdot A$,
(2) $(aA + bB) \cdot C = a(A \cdot C) + b(B \cdot C)$.

By (1), the map $(A, B) \to (A \cdot B)$ is symmetric; by (2), it is bilinear. The inner product is called positive-definite if $A \cdot A > 0$ for all $A \neq 0$. In that case the *norm* of the vector A is $|A| \equiv (A \cdot A)^{1/2}$. In relativity we deal with inner products that are indefinite: $A \cdot A$ has one sign for some vectors and another for others. In this case the norm, or magnitude, is often defined as $|A| \equiv |A \cdot A|^{1/2}$. Two vectors A and B are said to be orthogonal if and only if $A \cdot B = 0$.

It is often convenient to adopt a set of basis vectors $\{A_1, \ldots, A_n\}$ that are *orthonormal:* $A_i \cdot A_j = 0$ if $i \neq j$ and $|A_k| = 1$ for all k. This is not necessary, of course. The reader unfamiliar with nonorthogonal bases should try the following. In the two-dimensional Euclidean plane with Cartesian (orthogonal) coordinates x and y and associated Cartesian (orthonormal) basis vectors e_x and e_y, define A and B to be the vectors $A = 5e_x + e_y$, $B = 3e_y$. Express A and B as linear combinations of the nonorthogonal basis $\{e_1 = e_x, e_2 = e_y - e_x\}$. Notice that, although e_1 and e_x are the same, the 1 and x components of A and B are *not* the same.

Matrices

A matrix is an array of numbers. We shall only deal with square matrices, *e.g.*

$$\begin{pmatrix} 1 & 2 \\ 3 & 1 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 1 & 2 & 5 \\ -6 & 3 & 18 \\ 10^5 & 0 & 0 \end{pmatrix}.$$

The *dimension* of a matrix is the number of its rows (or columns). We denote the elements of a matrix by A_{ij}, where the value of i denotes the row and that of j denotes the column; for a 2 × 2 matrix we have

$$\mathbf{A} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}.$$

A column vector W is a set of numbers W_i, for example $\begin{pmatrix} W_1 \\ W_2 \end{pmatrix}$ in two dimensions. (Column vectors form a vector space in the usual way.) The following rule governs multiplication of a column vector by a matrix to give a column vector $V = \mathbf{A} \cdot W$:

$$\begin{pmatrix} V_1 \\ V_2 \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} W_1 \\ W_2 \end{pmatrix} = \begin{pmatrix} A_{11}W_1 + A_{12}W_2 \\ A_{21}W_1 + A_{22}W_2 \end{pmatrix}.$$

In index notation this is clearly

$$V_i = \sum_{j=1}^{2} A_{ij} W_j.$$

For n-dimensional matrices and vectors, this generalizes to

$$V_i = \sum_{j=1}^{n} A_{ij} W_j.$$

Notice that the sum is on the *second* index of **A**.

Matrices form a vector space themselves, with addition and multiplication by a number defined by:

$$\mathbf{A} + \mathbf{B} = \mathbf{C} \Rightarrow C_{ij} = A_{ij} + B_{ij}.$$
$$a\mathbf{A} = \mathbf{B} \Rightarrow B_{ij} = aA_{ij}.$$

For $n \times n$ matrices, the dimension of this vector space is n^2. A natural inner product may be defined on this space:

$$\mathbf{A} \cdot \mathbf{B} = \sum_{i,j} A_{ij} B_{ij}.$$

We can easily show that this is positive-definite. More important than the inner product, however, for our purposes, is *matrix multiplication*. (A vector space with multiplication is called an algebra, so we are now studying the matrix algebra.) For 2 × 2 matrices, the product is

$$\begin{aligned}
\mathbf{AB} = \mathbf{C} \Rightarrow \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}
&= \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \\
&= \begin{pmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{pmatrix}
\end{aligned}$$

In index notation this is

$$C_{ij} = \sum_{k=1}^{2} A_{ik} B_{kj}.$$

Generalizing to $n \times n$ matrices gives

$$C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}.$$

Notice that the index summed on is the second of A and the first of B. Multiplication is associative but not commutative; the identity is the matrix whose elements are δ_{ij}, the Kronecker delta symbol ($\delta_{ij} = 1$ if $i = j$, 0 otherwise).

The *determinant* of a 2×2 matrix is

$$\begin{aligned} \det \mathbf{A} &= \det \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \\ &= A_{11}A_{22} - A_{12}A_{21}. \end{aligned}$$

Given any $n \times n$ matrix $\mathbf{B}$ and an element B_{lm} (for fixed l and m), we call $\mathbf{S}_{lm}$ the $(n - 1) \times (n - 1)$ submatrix defined by excluding row l and column m from $\mathbf{B}$, and we call D_{lm} the determinant of $\mathbf{S}_{lm}$. For example, if $\mathbf{B}$ is the 3×3 matrix

$$\mathbf{B} = \begin{pmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \\ B_{31} & B_{32} & B_{33} \end{pmatrix},$$

then the submatrix $\mathbf{S}_{12}$ is the 2×2 matrix

$$\mathbf{S}_{12} = \begin{pmatrix} B_{21} & B_{23} \\ B_{31} & B_{33} \end{pmatrix}$$

and its determinant is

$$D_{12} = B_{21}B_{33} - B_{23}B_{31}.$$

Then the determinant of **B** is defined as

$$\det(\mathbf{B}) = \sum_{j=1}^{n} (-1)^{i+j} B_{ij} D_{ij} \quad \text{for any } i.$$

In this expression we sum only over j for fixed i. The result is independent of which i was chosen. This enables us to define the determinant of a 3 × 3 matrix in terms of that of a 2 × 2 matrix, and that of a 4 × 4 in terms of 3 × 3, and so on.

Because matrix multiplication is defined, it is possible to define the multiplicative inverse of a matrix, which is usually just called its inverse;

$$(\mathbf{B}^{-1})_{ij} = (-1)^{i+j} D_{ji} / \det(\mathbf{B})$$

The inverse is defined if and only if det (**B**) ≠ 0.

References

raham, R. & Marsden, J. E. (1978) *Foundations of Mechanics*, 2nd edn (Benjamin/ Cummings, Reading, Mass.).

ller, R., Bazin, M., & Schiffer, M. (1975) *Introduction to General Relativity*, 2nd edn (McGraw-Hill, New York).

cubierre, M. (2008) *Introduction to 3+1 Numerical Relativity* (Oxford University Press, Oxford).

len, S. W., Schmidt, R. W., & Fabian, A. C. (2002) *Mon. Not. Roy. Astron. Soc.*, **334**, L11.

ndersson, N., Kokkotas, K.D., & Schutz, B.F. (1999) *Astrophys. J.*, **510**, 846–853.

dréasson, H. (2005) 'The Einstein–Vlasov system/kinetic theory', *Living Rev. Relativity*, **8**, 2. URL: http://www.livingreviews.org/lrr-2005-2

nsorg, M. (2005) *Phys. Rev.*, **D72**, 024019.

mstrong, J. W. (2006) 'Low-frequency gravitational wave searches using spacecraft doppler Tracking', *Living Rev. Relativity*, **9**, 1. URL: http://www.livingreviews.org/lrr-2006-1

zeliès, H. (1966) *Relativistic Kinematics* (Pergamon Press, London). Translation of a 1955 French edn.

hby, N. (2003) 'Relativity in the global positioning system', *Living Rev. Relativity*, **6**, 1. URL: http://www.livingreviews.org/lrr-2003-1

shtekar, A. & Bojowald, M. (2006) *Class. Q. Grav.*, **23**, 391.

lian, R., Audouze, J., & Schramm, D. N. (1980) *Physical Cosmology*. Proceedings of 1979 Les Houches summer school (North-Holland, Amsterdam).

rrow, J. & Tipler, F. (1986) *The Anthropic Cosmological Principle* (Oxford University Press, Oxford).

ekenstein, J. D. (1973) *Phys. Rev. D*, **7**, 2333–2346.

ekenstein, J. D. (1974) *Phys. Rev. D*, **9**, 3292–3300.

rger, B. K. (2002) 'Numerical approaches to spacetime singularities', *Living Rev. Relativity*, **5**, 1. URL: http://www.livingreviews.org/lrr-2002-1

eyer, H. R. (2001) *Comm. Math. Phys.*, **221**, 659–676.

rrell, N. D. & Davies, P. C. W. (1984) *Quantum Fields in Curved Space* (Cambridge University Press, Cambridge).

shop, R. L. & Goldberg, S. I. (1981) *Tensor Analysis on Manifolds* (Dover, New York).

lair, D. (1991) *The Detection of Gravitational Waves* (Cambridge).

anchet, L. (2006) 'Gravitational radiation from post-Newtonian sources and inspiralling compact binaries', *Living Rev. Relativity*, **9**, 4. URL: http://www.livingreviews.org/lrr-2006-4

landford, R. D. & Znajek, R. L. (1977) *Mon. Not. Roy. astr. Soc.*, **179** 433–456.

hm, D. (2008) *The Special Theory of Relativity*, new edition with preface by J. Barrow (Routledge, New York).

jowald, M. (2005) 'Loop quantum cosmology', *Living Rev. Relativity*, **8**, 11. URL: http://www.livingreviews.org/lrr-2005-11

na, C. & Palenzuela-Luque, C. (2005) *Elements of Numerical Relativity* (Springer, Berlin).

riakoff, V., Ferguson, D. C., Haugan, M. P., Terzian, Y., & Teukolsky, S. A. (1982) *Astrophys. J.*, **261**, L101.

urne, D. E. & Kendall, P. C. (1992) *Vector Analysis and Cartesian Tensors* 3rd edn (CRC Press, New York).

renneman, L. & Reynolds, C. (2006) *Astrophys. J.*, **652** 1028.

uchdahl, H. A. (1959) *Phys. Rev.*, **116**, 1027.

ıchdahl, H. A. (1981) *Seventeen Simple Lectures on General Relativity Theory* (Wiley, New York).
utterworth, E. M. & Ipser, J. R. (1976) *Astrophys. J.*, **204**, 200.
ımpanelli, M., Lousto, C. O., Zlochower, Y., & Merritt, D. (2007) *Phys. Rev. Lett.*, **98**, 231102.
arroll, S. (2003) *Spacetime and Geometry* (Benjamin Cummings).
artan, E. (1923) *Ann. Ecole Norm. Sup.*, **40**, 325.
arter, B. (1969) *J. Math. Phys.*, **10**, 70.
ırter, B. (1973) In DeWitt, B. & De Witt, C. (eds.). *Black Holes* (Gordon and Breach, New York).
ırter, B. (1979) 'The general theory of the mechanical, electromagnetic, and thermodynamic properties of black holes'. In Hawking & Israel (1979), p. 294.
ıves, C. M., Thorne, K. S., Drever, R. W. P., Sandberg, V. D., & Zimmerman, M. (1980) *Rev. Mod. Phys.*, **52**, 341.
handrasekhar, S. (1957) *Stellar Structure* (Dover, New York).
ıandrasekhar, S. (1980) In Wayman, P., *Highlights of Astronomy* (Reidel, Dordrecht). Proceedings of the 1979 IAU meeting in Montreal, p. 45.
ıandrasekhar, S. (1983) *The Mathematical Theory of Black Holes* (Oxford University Press).
handrasekhar, S. (1987) *Truth and Beauty* (University of Chicago Press, Chicago).
ıoquet-Bruhat, Y., DeWitt-Morette, C., & Dillard-Bleick, M. (1977) *Analysis, Manifolds, and Physics* (North-Holland, Amsterdam).
hoquet-Bruhat, Y. & York, J. W. (1980) 'The Cauchy problem'. In Held (1980a), p. 99.
iufolini, I., Pavlis, E. C., & Peron, R. (2006) *New Astronomy*, **11**, 527–550.
omins, N. & Schutz, B. F. (1978) *Proc. Roy. Soc. A*, **364**, 211.
ıok, G. (2000) 'Initial Data for Numerical Relativity', *Living Rev. Relativity*, **3**, 5. URL: http://www.livingreviews.org/lrr-2000-5
ordes, J. M., Kramer, M., Lazio, T. J. W., *et al.* (2004) *New Astr. Rev.*, **48**, 1413–1438.
ımour, T. (1987) 'The problem of motion in Newtonian and Einsteinian gravity', in S. Hawking and W. Israel (eds.), *300 Years of Gravitation* (Cambridge), 128–198.
amour, T. & Dervelle, N. (1986) *Ann. Inst. Henri Poincare* **44**, 263–292.
eWitt, B. S. (1979) 'Quantum gravity: the new synthesis'. In Hawking & Israel (1979), p. 680.
icke, R. H. (1964) In Dewitt, B. & Dewitt, C. *Relativity, Groups and Topology* (Gordon and Breach, New York). Lectures at the 1963 Les Houches summer school.
ixon, W. G. (1978) *Special Relativity: The Foundation of Macroscopic Physics* (Cambridge University Press, Cambridge).
hlers, J., and Kundt, W. (1962), in Witten, L., ed., *Gravitation: An Introduction to Current Research* (Wiley, New York), pp. 49–101.
ermi, E. (1956) *Thermodynamics* (Dover, New York).
lippenko, A. (2008) *Dark Energy and Supernovae* (Princeton University Press,

Princeton).
ench, A. P. (1968) *Special Relativity* (W. W. Norton, New York).
iedman, J. L. (1978) *Commun. Math. Phys.*, **63**, 243.
iedman, J. L. & Schutz, B. F. (1978) *Astrophys. J.*, **222**, 281.
itamase, T. (1983) *Phys. Rev. D*, **28**, 2373.
itamase, T. & Itoh, Y. (2007), *Living Rev. Relativity*, **10**, 2. URL: http://www.livingreviews.org/lrr-2007-2
itamase, T. & Schutz, B. F. (1983) *Phys. Rev. D*, **28**, 2363.
enzel, R. & Karas, V. (2007) in V. Karas and G. Matt (eds.) *Black Holes* (Cambridge University Press, Cambridge), pp. 173–180.
eroch, R. (1978) *General Relativity from A to B* (University of Chicago Press).
eroch, R. & Horowitz, G. T. (1979) 'Global structure of spacetimes'. In Hawking & Israel (1979), p. 212.
lendenning, N. K. (2007) *Special and General Relativity: With Applications to White Dwarfs, Neutron Stars and Black Holes* (Springer, New York).
ron, O. & Hervik, S. (2007) *Einstein's General Theory of Relativity: With Modern Applications in Cosmology* (Springer, New York).
uth, A. (1981) *Phys. Rev. D*, **23**, 347.
agedorn, R. (1963) *Relativistic Kinematics* (Benjamin, New York).
ansen, C. J., Kawaler, S. D., & Trimble, V. (2004) *Stellar Interiors – Physical Principles, Structure, and Evolution* (Springer-Verlag, New York).
arrison, B. K., Thorne, K. S., Wakano, M., & Wheeler, J. A. (1965) *Gravitation Theory and Gravitational Collapse* (University of Chicago Press).
artle, J. B. (2003) *Gravity: An Introduction to Einstein's General Relativity* (Benjamin Cummings).
awking, S. W. (1972) *Commun. Math. Phys.*, **25**, 152.
awking, S. W. (1975) *Commun. Math. Phys.*, **43**, 199.
awking, S. W. & Ellis, G. F. R. (1973) *The Large-Scale Structure of Space–Time* (Cambridge University Press, Cambridge).
awking, S. W. & Israel, W. (1979) *General Relativity: An Einstein Centenary Survey* (Cambridge University Press, Cambridge).
eidmann, J. (1980) *Relativistic Cosmology* (Springer, Berlin).
eld, A. (1980a) *General Relativity and Gravitation* (Plenum, London), vol. 1.
eld, A. (1980b) *General Relativity and Gravitation* (Plenum, London), vol. 2.
ermann, R. (1968) *Differential Geometry and the Calculus of Variations* (Academic Press, New York).
icks, N. J. (1965) *Notes on Differential Geometry* (D. Van Nostrand, New York).
eusler, M. (1998) 'Stationary black holes: uniqueness and beyond', *Living Rev. Relativity*, **1**, 6. URL: http://www.livingreviews.org/lrr-1998-6

obson, M. P., Efstathiou, G. P., & Lasenby, A. N. (2006) *General Relativity: An Introduction for Physicists* (Cambridge University Press, Cambridge).

offmann, B. (1983) *Relativity and Its Roots* (Dover, New York).

öflich, P., Kumar, P., & Wheeler, J. C. (2004) *Cosmic Explosions in Three Dimensions: Asymmetries in Supernovae and Gamma-Ray Bursts* (Cambridge University Press, Cambridge).

ough, J. L. and Rowan, S. (2000) 'Gravitational wave detection by interferometry (ground and space)', *Living Rev. Relativity*, **3**, 3. URL: http://www.livingreviews.org/lrr-2000-3

ulse, R. A. and Taylor, J. H. (1975) 'Discovery of a pulsar in a binary system', *Astrophys. J.*, **195**, L51–L53.

ham, C. J. (1999) *Modern Differential Geometry for Physicists* (World Scientific, Singapore).

rael, W. & Stewart, J. (1980) 'Progress in relativistic thermodynamics and electrodynamics of continuous media'. In Held (1980), p. 491.

ackson, J. D. (1975) *Classical Electrodynamics* (Wiley, New York).

ilmister, C. W. (1970) *Special Theory of Relativity* (Pergamon, Oxford).

nop, R.A. *et al*. (2003) *Astrophys. J.*, **598**, 102.

obayashi, S. & Nomizu, K. (1963) *Foundations of Differential Geometry* (Interscience, New York), vol. I.

obayashi, S. & Nomizu, K. (1969) *Foundations of Differential Geometry* (Interscience, New York), vol II.

okkotas, K. D. & Schmidt, B. (1999) 'Quasi-normal modes of stars and black holes', *Living Rev. Relativity*, **2**, 2. URL: http://www.livingreviews.org/lrr-1999-2

omossa, S., Zhou, H., & Lu, H. (2008) 'A recoiling supermassive black hole in the quasar SDSS J092712.65+294344.0?', *Astrophys. J. Lett.*, **678**, L81–L84.

ramer, M. *et al.* (2006) *Science*, **314**, 97–102.

andau, L. D. & Lifshitz, E. M. (1959) *Fluid Mechanics* (Pergamon, Oxford).

andau, L. D. & Lifshitz, E. M. (1962) *The Classical Theory of Fields* (Pergamon, Oxford).

attimer, J. M. & Prakash, M. (2000) 'Nuclear matter and its role in supernovae, neutron stars and compact object binary mergers', *Phys. Rep.*, **333** 121–146.

iang, E. P. T. & Sachs, R. K. (1980) 'Cosmology'. In Held (1980b), p. 329.

iddle, A. (2003) *An Introduction to Modern Cosmology*, 2nd. edn (Wiley, New York).

ghtman, A. P., Press, W. H., Price, R. H., & Teukolsky, S. A. (1975) *Problem Book in Relativity and Gravitation* (Princeton University Press, Princeton).

inde, A. (1982) *Phys. Lett.*, **108B**, 389.

orimer, D. R. (2008) 'Binary and millisecond pulsars', *Living Rev. Relativity*, **11**, 8. URL: http://www.livingreviews.org/lrr-2008-8

orimer, D. R. & Kramer, M. (2004) *Handbook of Pulsar Astronomy* (Cambridge University Press, Cambridge).

ovelock, D. & Rund, H. (1990) *Tensors, Differential Forms, and Variational Principles* (Dover, New York).

yne, A.G. & Smith, F. G. (1998) *Pulsar Astronomy* (Cambridge University Press, Cambridge).

aartens, R. (2004) 'Brane–world gravity', *Living Rev. Relativity*, **7**, 7. URL: http://www.livingreviews.org/lrr-2004-7

acCallum, M. A. H. (1979) 'Anisotropic and inhomogeneous relativistic cosmologies'. In Hawking & Israel (1979), p. 533.

athews, J. & Walker, R. L. (1965) *Mathematical Methods of Physics* (Benjamin, New York).

IcCrea, W. H. (1979), *Q. J. Roy. Astron. Soc.*, **20**, 251.

Iarder, L. (1971) *Time and the Space-Traveller* (George Allen & Unwin, London).

ermin, N. D. (1989) *Space and Time in Special Relativity* (Waveland Press, Illinois).

erritt, D. and Milosavljevic, M. (2005) 'Massive black hole binary evolution', *Living Rev. Relativity*, **8**, 8. URL: http://www.livingreviews.org/lrr-2005-8

iller, M. C. & Colbert, E. J. M. (2004) *Int. J. Mod. Phys.*, **D13**, 1–64.

isner, C. W., Thorne, K. S., & Wheeler, J. A. (1973) *Gravitation* (Freeman, San Francisco).

øller, C. (1972) *The Theory of Relativity* (Clarendon Press, Oxford).

ukhanov, V. (2005) *Physical Foundations of Cosmology* (Cambridge University Press, Cambridge).

is, A. (1982) *'Subtle is the Lord . . .', The Science and the Life of Albert Einstein* (Oxford University Press, Oxford).

ebles, P. J. E. (1980) *Large Scale Structure of the Universe* (Princeton University Press).

enrose, R. (1969), *Rev. Nuovo Cimento* **1**, 252.

enrose, R. (1979) In Hawking & Israel (1979), 581.

enzias, A. A. (1979) *Rev. Mod. Phys.*, **51**, 425.

rlick, V. (2004) 'Gravitational lensing from a spacetime perspective', *Living Rev. Relativity*, **7**, 9. URL: http://www.livingreviews.org/lrr-2004-9

rlmutter, S., Aldering, G., Goldhaber, G., Knop, R.A., Nugent, P., Castro, P.G., Deustua, S., Fabbro, S., Goobar, A., Groom, D., Hook, I.M., Kim, A.G., Kim, M.Y., Lee, J.C., Nunes, N.J., Pain, R., Pennypacker, C.R., Quimby, R., Lidman, C., Ellis, R.S., Irwin, M., McMahon, R.G., Ruiz-Lapuente, P., Walton, N., Schaefer, B., Boyle, B.J., Filippenko, A.V., Matheson, T., Fruchter, A.S., Panagia, N., Newberg, H.J.M., and Couch, W.J., (1999), *Astrophys. J.*, **517**, 565–586.

eters, P. C. (1964) *Phys. Rev.*, **B 136**, 1224.

ound, R. V. & Rebka, G. A. (1960) *Phys. Rev. Lett.*, **4**, 337.

ound, R. V. & Snider, J. L. (1965) *Phys. Rev.*, **B 140**, 788.

ice, R. H. (1972a) *Phys. Rev. D*, **5**, 2419.

ice, R. H. (1972b) *Phys. Rev. D*, **5**, 2439.

ndall, A. D. (2005) 'Theorems on existence and global dynamics for the Einstein equations', *Living Rev. Relativity*, **8**, 6. URL: http://www.livingreviews.org/lrr-2005-6

nn, J. ed. (2007) *The Genesis of General Relativity: Sources and Interpretations* (Springer, 4 vols).

ess, A.G., Filippenko, A.V., Challis, P., Clocchiatti, A., Diercks, A., Garnavich, P.M., Gilliland, R.L., Hogan, C.J., Jha, S., Kirshner, R.P., Leibundgut, B., Phillips, M.M., Reiss, D., Schmidt, B.P., Schommer, R.A., Smith, R.C., Spyromilio, J., Stubbs, C., Suntzeff, N.B., and Tonry, J. (1998) *Astron. J.*, **116**, 1009–1038.

ndler, W. (1991) *Introduction to Special Relativity*, 2nd edn. (Oxford University Press, Oxford).

ndler, W. (2006) *Relativity: Special, General, Cosmological*, 2nd edn. (Oxford University Press, Oxford).

obinson, D. C. (1975) *Phys. Rev. Lett.*, **34**, 905.

ovelli, C. (2004) *Quantum Gravity* (Cambridge University Press, Cambridge).

an, M. P. & Shepley, L. C. (1975) *Homogeneous Relativistic Cosmologies* (Princeton University Press).

ulson, P. (1994) *Funamentals of Interferometric Gravitational Wave Detectors* (World Scientific).

hild, A. (1967) In Ehlers, J. (ed.), *Relativity Theory and Astrophysics: I. Relativity and Cosmology* (American Math. Soc., Providence, R.I.).

chneider, P. (2006) *Extragalactic Astronomy and Cosmology* (Springer-Verlag, Berlin).

hneider, P., Ehlers, J., Falco, E. (1992) *Gravitational Lenses* (Springer-Verlag, Berlin).

:houten, J. A. (1990) *Tensor Analysis for Physicists* (Dover, New York).

:hrödinger, E. (1950) *SpaceTime Structure* (Cambridge University Press, Cambridge).

:hutz, B. F. (1980a) *Phys. Rev.*, **D 22**, 249.

hutz, B. F. (1980b) *Geometrical Methods of Mathematical Physics* (Cambridge University Press, Cambridge).

:hutz, B. F. (1984) *Am. J. Phys.*, **52**, 412.

:hutz, B. F. (1986) *Nature*, **323**, 310–311.

hutz, B. F. (2003) *Gravity from the Ground Up* (Cambridge University Press, Cambridge).

hutz, B. F. & Ricci, F. (2001) in I. Ciufolini, *et al*. (eds.) *Gravitational Waves* (Inst. of Phys., Bristol), pp. 11–83.

hwarz, P. M. & Schwarz, J. H. (2004) *Special Relativity: From Einstein to Strings* (Cambridge University Press, Cambridge).

apiro, I. I. (1980) 'Experimental tests of the general theory of relativity'. In Held (1980b), p. 469.

apiro, S. L. & Teukolsky, S. A. (1983) *Black Holes, White Dwarfs, and Neutron Stars* (Wiley, New York).

narr, L. L. (ed.) (1979) *Sources of Gravitational Radiation* (Cambridge University Press, Cambridge).

pergel, D. N. *et al* (2003) *Astrophys. J. Suppl.*, **148** 175.

ivak, M. (1979) *A Comprehensive Introduction to Differential Geometry*, (Publish or Perish, Boston), vols. 1–5.

airs, I. H. (2003) 'Testing general relativity with pulsar timing', *Living Rev. Relativity*, **6**, 5. URL: http://www.livingreviews.org/lrr-2003-5

:arobinsky, A. (1980) *Phys. Lett.*, **91B**, 99.

ephani, H. (2004) *Relativity: An Introduction to Special and General Relativity*, 3rd edn (Cambridge University Press, Cambridge).

ephani, H., Kramer, D., MacCallum, M., Hoenselaers, C., and Herlt, E. (2003), *Exact Solutions of Einstein's Field Equations*, 2nd edn (Cambridge University Press).

ergioulas, N. (2003) 'Rotating Stars in Relativity', *Living Rev. Relativity* **6**, 3. URL: http://www.livingreviews.org/lrr-2003-3

ynge, J. L. (1960) *Relativity: The General Theory* (North-Holland, Amsterdam).

ynge, J. L. (1965) *Relativity: The Special Theory* (North-Holland, Amsterdam).

yler, R. J. (1994) *The Stars: their Structure and Evolution*, 2nd edn (Cambridge University Press, Cambridge).

aylor, E. F. & Wheeler, J. A. (1966) *Spacetime Physics* (Freeman, San Francisco).

aylor, J. H. & Weisberg, J. M. (1982) *Astrophys. J.*, **253**, 908.

erletskii, Y. P. (1968) *Paradoxes in the Theory of Relativity* (Plenum, New York).

eukolsky, S. A. (1972) *Phys. Rev. Lett.*, **29**, 114.

iiemann, T. (2007) *Modern Canonical Quantum General Relativity* (Cambridge University Press, Cambridge).

iorne, K. S., Caves, C. M., Sandberg, V. D., & Zimmerman, M. (1979) 'The quantum limit for gravitational-wave detectors and methods of circumventing it'. In Smarr (1979a), p. 49.

pler, F. J., Clarke, C. J. S., & Ellis, G. F. R. (1980), 'Singularities and horizons – a review article'. In Held (1980b), p. 97.

'ald, R. M. (1984) *General Relativity* (Chicago University Press, Chicago).

ald, R. M. (1994) *Quantum Field Theory in Curved Spacetime and Black Hole Thermodynamics* (Chicago University Press, Chicago).

ambsganss, J. (1998) 'Gravitational lensing in astronomy', *Living Rev. Relativity*, **1**, 12. URL: http://www.livingreviews.org/lrr-1998-12

'ard, W. R. (1970) *Astrophys. J.*, **162**, 345.

eber, J. (1961) *General Relativity and Gravitational Waves* (Interscience, New York).

'einberg, S. (1972) *Gravitation and Cosmology* (Wiley, New York).

eisberg, J. M. & Taylor, J. H. (2005), in F. A. Rasio and I. H. Stairs (eds.), *Binary Radio Pulsars* (Astr. Soc. Pacific, San Francisco), pp. 25–32.

'hiting, B. F. (1989) *J. Math. Phys.*, **30**, 1301–1305.

'iedemann, H. (2007) *Particle Accelerator Physics*, 3rd edn (Springer, Berlin).

'ill, C. M. (1974) *Astrophys. J.*, **191**, 521.

'ill, C. M. (1975) *Astrophys. J.*, **196**, 41.

'ill, C. M. (1993) *Theory and Experiment in Gravitational Physics* (Cambridge University Press, 2nd edn.).

'ill, C. M. (2006) 'The confrontation between general relativity and experiment', *Living Rev. Relativity*, **9**, 3. URL: http://www.livingreviews.org/lrr-2006-3

'illiams, L. P. (1968) *Relativity Theory: Its Origins and Impact on Modern Thought* (Wiley, New York).

'ilson, R. W. (1979) *Rev. Mod. Phys.*, **51**, 433.

'oodhouse, N. M. J. (2003) *Special Relativity* (Springer, London).

'oodhouse, N. M. J. (2007) *General Relativity* (Springer, London).

oosley, S. E. & Bloom, J. S. (2006) 'The Supernova Gamma-Ray Burst Connection', *Ann. Rev. Astron. Astroph.*, **44** 507–556.

ıno, K. (1955) *The Theory of Lie Derivatives and its Applications* (North-Holland, Amsterdam).

ıo, W.-M. *et al.* (2006) *J. Phys. G.*, **33**, 1. This is kept up to date on the Particle Data website, http://pdg.lbl.gov/2006/reviews/contents_sports.html

ı'dovich, Ya. B. & Novikov, I. D. (1971) In Thorne, K. S & Arnett, W. D. (eds.), *Stars and Relativity (Relativistic Astrophysics Vol. I)* (University of Chicago Press).

Index